Treball final de grau

GRAU DE MATEMÀTIQUES

Facultat de Matemàtiques i Informàtica
Universitat de Barcelona

Uso del análisis de clústeres para determinar las características de los mercados financieros

Autor: David Portabella de Pedro

Director: Dr. Josep Vives Santa-Eulalia
Dr. Jose Bonifacio Saez Madrid
Realitzat a: Departament de Probabilitat, Lógica i Estadistica
Departament de Matemàtica Econòmica, Financera i Actuarial

Barcelona, 27 de junio de 2018

1. Resumen

Este trabajo aplica el análisis de clústeres a los mercados financieros para sacar conclusiones tanto de los conjuntos de empresas con comportamientos similares como sobre la composición de dichos grupos, buscando ver si empresas de diferentes mercados financieros se agrupan juntas o es independiente el factor mercado.

Comienza desarrollando el concepto de análisis de clúster y observando las variables del mercado financiero que se analizarán. A continuación concluye que uno de los mayores problemas del análisis de clústeres es la existencia de variables con distintas ponderaciones y se diseña un método llamado "método de la partición más estable" para solventar el problema. Finalmente se lleva a la práctica el método programándolo en R y se obtienen las conclusiones de los mercados financieros.

2. Abstract

This paper applies cluster analysis to financial markets in order to draw conclusions both about the groups of companies with similar behaviour and about the composition of these groups, seeking to understand whether companies from different financial markets are grouped together or whether the market factor is independent.

It begins by developing the concept of cluster analysis and looking at the financial market variables to be analyzed. Then, it concludes that one of the biggest issues with cluster analysis is the existence of variables with different weights and it designs a method called "the most stable partition method" to solve the problem. Finally, the method is implemented by programming it in R and the conclusions of the financial markets are obtained.

3. Agradecimientos

Este trabajo no es solo mío sino también de toda la gente que ha contribuido a él de forma directa o indirecta. En primer lugar hay que destacar la contribución de ambos tutores de este trabajo.

Por un lado Josep Vives por su dedicación. Habitualmente no agradecemos suficiente el tiempo que se nos dedica y solemos dar por sentado que debe ser así. Josep, quien a pesar de estar cumpliendo las obligaciones básicas de la universidad y otras adquiridas, ha demostrado ser una persona dispuesta a dar esa dedicación y a lidiar con una cantidad notable de trabajos de distintos alumnos, no solo con él presente texto, si así podía contribuir a ayudar a los alumnos de la universidad. Creo que ahí esta parte de la esencia de ser un buen profesor.

Por otro lado a José Sáez por sus conocimientos sobre los mercados financieros y por ayudarme a ver que todas las variables se pueden ver desde distintos puntos de vista y calcular muchas variantes. Hice este trabajo para aprender más sobre bolsa y el ha sido la principal razón de que aprendiese algo nuevo sobre los mercados financieros.

Además quiero agradecer a mi familia y amigos, quienes siempre han ofrecido su apoyo y su comprensión y en particular a Nuria porque sin sus consejos el documento seria la mitad de lo que es.

Índice

4. Introducción

El análisis de clústeres permite analizar un conjunto de observaciones para determinar similitudes y diferencias entre ellas. Se aplicará el análisis de clústeres a la información financiera de las empresas de tres mercados distintos: Ibex, Dow Jones y Euro Stoxx. Tras analizar las empresas a nivel global se pretende crear asociaciones entre ellas para determinar dos hechos:

Por un lado de cara a un inversor que actúe a corto plazo, se pretende determinar diferentes conjuntos de empresas en el panorama internacional, facilitando la tarea de diversificar la cartera al invertir en diferentes conjuntos que presenten diferentes comportamientos para maximizar el rendimiento de su inversión.

Por el otro lado determinar si la distribución de empresas en clústeres se realiza de forma homogénea, es decir, si a nivel global, empresas de diferentes países se comportan igual o si, por lo contrario, cada mercado define en su mayoría un clúster, significando que existe una tendencia diferente por mercado.

Para ello el trabajo consta de tres partes:

En una primera parte se introduce el análisis de clústeres a la vez que se analizan sus propiedades, puntos fuertes y débiles. Así mismo se seleccionan las variables que se analizarán para el estudio de los mercados financieros. Se concluye que el mayor punto débil de los análisis de clústeres es la falta de criterios para ponderar variables que pueden tener diferente importancia para el estudio.

En una segunda parte se profundiza sobre este problema y se analiza y diseña una variante de los métodos existentes con el objetivo de solventar este inconveniente. El método resultante se denominará método de la partición más estable.

En la tercera y última parte se programa el método definido anteriormente, solventando los distintos temas que surgen durante el proceso de llevarlo a la práctica. Finalmente, se realiza el análisis de las distintas bolsas aplicando el método de la partición más estable para obtener las conclusiones del estudio.

5. Análisis de clústeres

El análisis de clústeres es la rama de las matemáticas que se encarga de analizar un conjunto de observaciones, definir unos criterios de clasificación y obtener una partición de los datos cuyos conjuntos sean lo más homogéneos posible y a la vez estos conjuntos sean lo más diferentes entre ellos según los criterios determinados previamente.

Denominaremos al conjunto de todas las observaciones realizadas X. Cada una de estas observaciones puede ser tanto una única variable como un vector de varias dimensiones. En este último caso representaremos cada observación de la forma $(x_1, ..., x_n) = x \in X$.

Un ejemplo de esto se puede ver si nos disponemos a analizar un conjunto de personas. En este caso, cada observación seria una persona del conjunto, de cada una ellas se podría analizar los datos relativos a su altura, su peso y su edad entre otros. De esta manera, tras aplicar un método de análisis de clústeres, como les que se explicarán a continuación, se obtendrían una partición del conjunto de personas. Cada subconjunto de la partición seria un grupo de personas con características similares, por ejemplo podría darse un subconjunto que se caracterizase por estar compuesto por personas que fuesen de media edad y altas.

Las funciones del análisis de clústeres son identificar conjuntos para poder nombrarlos, resumir datos, predecir o obtener explicaciones. Los elementos se agrupan de manera que diferencias sutiles pueden volverse más aparentes al separar los elementos en diferentes clústeres, además permiten suponer propiedades de los elementos ya que si ciertos objetos de un clúster cumplen una propiedad, es posible que otros objetos del mismo clúster cumplan la misma propiedad.

5.1. Tipologías de variables:

En el momento de realizar análisis de clústeres se pueden utilizar datos muy diferentes entre sí. Para un mejor tratamiento de las variables es recomendable clasificar cada una de ellas según su tipología. Hartigan [1] nos indica que una forma de clasificar las diferentes variables es:

a) Cuentas: Sin escala arbitraria (ej.: número de ojos de una hormiga)

b) Ratios: Determinados por su tamaño respecto a un volumen estándar (ej.: cantidad de agua en un vaso)

c) Medida de intervalo: Determinados respecto a un punto de referencia arbitrario y en términos de unas unidades. (ej.: altura de una montaña en metros respecto al nivel del mar)

d) Escala ordinal: Los objetos pueden ser ordenados aunque la diferencias entre valores no es significativa. (ej.: posiciones de una carrera, a priori no sabemos la diferencia de tiempo del primero al segundo. Y aun de saberla, no daría ninguna información sobre la diferencia de tiempo entre el segundo y el tercero)

e) Escala nominal: Los elementos se pueden poner en categorías para las cuales no se ha definido siquiera un orden. (ej.: religión que práctica una persona)

Como se ha dicho, el principal objetivo cuando se aplica un método de clustering es obtener una partición de los datos cuyos conjuntos sean lo más homogéneos posibles. Para determinar esta homogeneidad, Arratia [2] nos indica que se realiza a través de una función que se debe minimizar.

Esta función suele estar expresado en términos de distancia o en ocasiones se aplican también métodos basados en coeficientes de correlación. En los casos que las variables observadas no sean numerales sino categorías, se suelen usar criterios basados en la posesión o falta de los atributos.

5.2. Distancias

Se denomina distancia a una función

$$d(\ ,\ ) : X \times X \to \mathbb{R}^+$$

tal que para todo $x, y, z \in X$ verifica las siguientes propiedades:

a)$d(x, x) = 0$

b)$d(x, y) = d(y, x)$

c)$d(x, y) \leq d(x, z) + d(z, y)$

Estas funciones se utilizan intuitivamente para medir la similitud entre dos observaciones, calculándolas de manera que una distancia elevada indica disimilaridad y una distancia reducida indique similitud.

Distancia euclidea:

Existe una gran variedad de distancias con diferentes propiedades, una de las más utilizadas es la distancia euclidea. Para dos elementos $x, y \in X$ $x = (x_1, ..., x_n),\ y = (y_1, ..., y_n)$ tiene el siguiente aspecto:

$$D(x, y) = \sqrt{\sum_{k=0}^{n} (x_k - y_k)^2}$$

Uno de los factores más importantes de su gran popularidad es que coincide con el concepto común de distancia, lo cual la hace más simple de aplicar.

Por otro lado, el uso de esta distancia también presenta ciertos aspectos no tan positivos. Sus principales inconvenientes son:

-Es sensible a las unidades de las variables: en caso de que haya variables de diferente tamaño, las unidades de mayor medida se ven más reflejadas en esta distancia que las de menor tamaño. Un ejemplo de este problema sería el realizar clústeres con observaciones sobre pesos y alturas.

-Si las variables están correlacionadas, la distancia reflejara más de una vez el mismo impacto. Se puede solucionar el problema seleccionando las causas originales de dichas variables, las cuales sí que estarían incorrelacionadas.

La distancia euclidea es de gran utilidad a la hora de realizar estudios con variables que estén incorrelacionadas y se presenten en unidades similares.

Distancia euclidea ponderada:

Existe una variante de la distancia euclidea diseñada para lidiar con el problema de la sensibilidad frente a las unidades, esta variante es la llamada distancia euclidea ponderada. Consiste en aplicar una función de ponderación a las distintas variables con la finalidad de hacerlas comparables. También se utiliza para resaltar o disminuir la importancia de ciertas variables en el estudio. Tras definir unos valores de ponderación $w() : \{1, ..., n\} \to \mathbb{R}^+$, la fórmula de la distancia quedaría:

$$D_w(x, y) = \sqrt{\sum_{k=0}^{n} w_k(x_k - y_k)^2}$$

Existen una gran variedad de métodos para seleccionar estas ponderaciones, algunos de ellos son:

-Subjetivo: El analista asigna a las variables ponderaciones que considera que representan mejor la realidad, bajo su propio criterio.

- Error de medida: Los pesos se seleccionan de forma inversamente proporcional al error de medida. Por ejemplo, si se estudian datos sobre personas en diferentes poblaciones, las ponderaciones se calcularían a partir de la media de varianzas para cada población. Este tipo de ponderación requiere de la formación de clústeres previos al análisis, lo cual podría no ser posible de efectuar.

- Misma escala de varianza: Los pesos corresponden a la inversa de la varianza, por lo tanto estos pesos son elegidos a partir de los datos que se estudian. Si, por ejemplo, se estudiasen un conjunto de personas y las variables en las que se basa el estudio son la edad y la altura, para equilibrar estos números tan dispares, en primer lugar se deberían calcular las varianzas. Supongamos que los resultado obtenidos son $VAR(Edad) = 4,25\ años^2$, $Var(ALT) = 9\ cm^2$, entonces la distancia resultante se calcularía como:

$$d(x, y) = \sqrt{(\frac{1}{4,25}(x_1 - y_1)^2 + \frac{1}{9}(x_2 - y_2)^2)}$$

Bajo esta medida las variables son invariables a cambios en las unidades de medida y todas hacen la misma contribución media a la distancia.

Distancia Mahalanobis

Otra metodologia de considerar la distancia es la distancia de Mahalanobis [6]

que viene determinada por

$$d^2(x,y) = ((x_1,...,x_n) - (y_1,...,y_n))V^{-1}\begin{pmatrix} x_1 & - & y_1 \\ & \vdots & \\ x_n & - & y_n \end{pmatrix}$$

Donde V^{-1} representa la inversa de la matriz de varianzas. Esta distancia tiene dos ventajas respecto a la distancia euclidea. La primera es que la distancia no depende de las unidades de medida en que se hayan estudiado las variables.

Para verlo consideremos el vector $W = (w_1, ..., w_n)$ que representa las variables originales y consideramos su transformación lineal a nuevas variables $W^* = (w^*_1, ..., w^*_n)$. Esta transformación lineal se identifica con la matriz C. Su relación viene determinada por la ecuación $W^* = CW$

Asimismo la matriz de varianzas para las variables W^* ahora es:

$$V^* = C^{-1}VC$$

La distancia de Mahanalanobis sobre las variables transformadas será

$$d^2(x^*, y^*) = (x^* - y^*)^t V^{*-1}(x^* - y^*) =$$
$$= (x-y)^t C(C^{-1}V^{-1}C)C^{-1}(x-y) = (x-y)^t V^{-1}(x-y)$$

que coincide con la distancia de Mahalanobis respecto las variables originales.

Por otro lado, al utilizar la inversa de la matiz de varianzas,la distancia de Mahalanobis corrige las correlaciones entre variables y de esta forma se reduce la redundancia.

5.3. Tipologías de métodos de análisis

Para la realización de un análisis de clústeres, además de seleccionar las variables a estudiar y con que distancia se medirán las diferencias entre variables, se debe seleccionar que método de análisis de clústeres se aplicará.

Existen diversos métodos de análisis cada uno de ellos diseñado en función de diferentes criterios. Es importante por lo tanto seleccionar un método adecuado al estudio que se pretende realizar puesto que métodos diferentes pueden llevar a resultados diferentes.

Hay dos criterios principales que se suelen seguir a la hora de definir los clústeres resultantes.

Por un lado se busca la cohesión de los clústeres resultantes. Esto es, se busca reducir la distancia media entre elementos dentro del mismo clúster de forma que se pueda afirmar que los elementos de este son muy similares entre si.

Por otro lado se busca la diferencia entre clústeres. En general esto se obtiene mediante criterios que maximizan la distancia que habrá entre los clústeres resultantes. De esta forma se puede afirmar que dos elementos de dos clústeres diferentes presentan unas características lo bastante dispares como para que merezca la pena clasificarlos de forma diferente

Un factor crítico a tener en cuenta a la hora de seleccionar el método de análisis es el algortimo que sigue este método. Los diferentes algoritmos influyen altamente los clústeres resultantes.

En función del algoritmo de análisis que se realice se pueden clasificar en dos grandes categorías de métodos de clustering, métodos jerárquicos y métodos no jerárquicos. [5]

Métodos de análisis jerárquicos

Los métodos jerárquicos incluyen todos esos métodos de clustering en los que se ejecutan sucesivas particiones cada una dando lugar a diferentes niveles de clasificación. En esta clase de métodos los conjuntos de una clasificación se unen para dar lugar a los conjuntos de otras clasificaciones más genéricas.

En función de la partición inicial que se utilice se pueden definir los métodos jerárquicos aglomerativos y los métodos jerárquicos divisivos.

a) Métodos jerárquicos aglomerativos: estos métodos comienzan con una partición en la cual cada elemento forma su propio clúster, de manera que hay tantos clústeres como objetos iniciales. En cada uno de los sucesivos pasos se agrupan objetos similares, dando lugar a particiones con un número menor de clúster hasta que en el último paso se dispone de un único clúster con todos los elementos.

b) Métodos jerárquicos divisivos: estos métodos funcionan a la inversa. Se inicia el algoritmo con una partición que consta de un único clúster conteniendo todos los elementos. En los sucesivos pasos se divide de forma que se separan los elementos más diferentes en nuevos clústeres hasta terminar con una partición en que cada elemento tiene su propio clúster.

Una de las ventajas de estos métodos de clustering es que permiten la creación de gráficos que ilustren las agrupaciones etapa a etapa. Estos gráficos en forma de árbol se denominan dendogramas. Su utilidad radica en que ofrecen una representación de las conclusiones del análisis de clústeres de forma intuitiva y visual. Un ejemplo de dendograma es la figura 1.

Los métodos jerárquicos presentan una serie de inconvenientes:

— Las fuentes de error y variación no entran en consideración con los métodos jerárquicos. Esto implica una gran sensibilidad a observaciones anómalas o outliers.

— Si un objeto se ha colocado erróneamente en un grupo al principio del proceso, ya no se puede arreglar en una etapa posterior.

Una técnica para tratar de mitigar estos inconvenientes consiste en usar varias distancias y observar si se mantienen los mismos clústeres. De esta forma se comprueba la existencia de grupos similares.

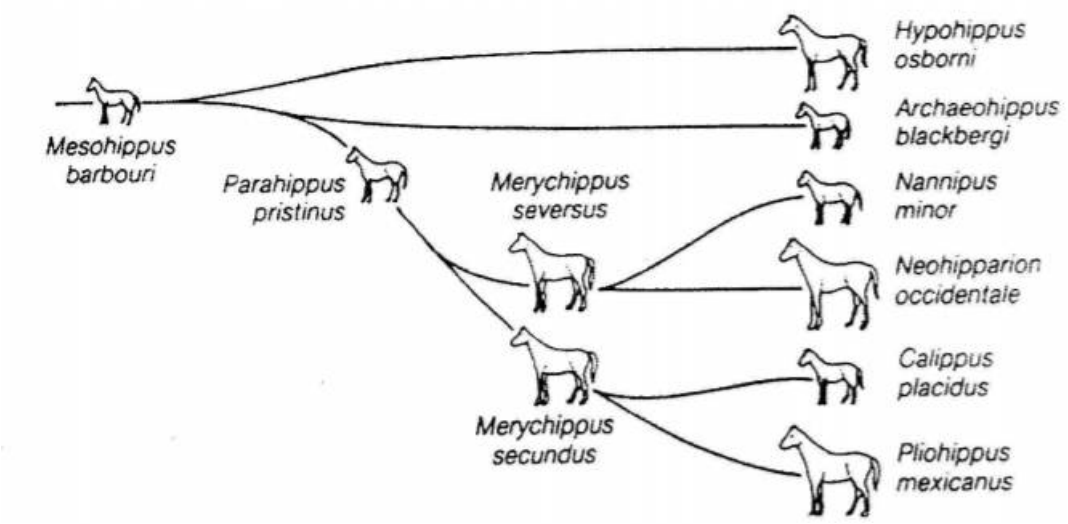


Figura 1: Dendograma

Métodos de análisis no jerárquicos

Los métodos no jerárquicos están diseñados para clasificar individuos en una clasificación de K clústeres, donde K se especifica a priori o bien se determina como una parte del proceso.

Los algoritmos utilizados para encontrar la partición en estos métodos en general se inician a partir de una partición inicial elegida de forma más o menos arbitraria. A partir de esta se intercambian elementos entre los clústeres de esta partición con la finalidad de obtener una partición más homogénea.

Los algoritmos continúan efectuando variaciones entre los clústeres hasta que alcanzan un óptimo local.

Los métodos no jerárquicos presentan la ventaja que no requieren que se guarde la información sobre todas las iteraciones anteriores y en muchos casos necesitan calcular un menor número de cálculos de distancias, lo que permitiría computar un mayor volumen de datos que con los métodos jerárquicos.

Uno de los métodos no jerárquicos más extendidos es el método de las K-medias, también conocido por su nombre en inglés, K-Means.

Método de K-Means

Este método permite asignar los elementos a un conjunto de k clústeres de manera que se minimiza la distancia media entre elementos de cada clúster. El algoritmo consiste a grades rasgos en asignar a cada observación el clúster cuyo centroide está más cercano a dicha observación. En general, la distancia empleada es la euclídea.

El método en detalle sería:

1. Se toman según el criterio del analista, k clústeres iniciales. En su defecto se pude optar por seleccionar únicamente k puntos iniciales e ir actualizando los centroides de los clústeres a medida que se añadan puntos.

2. Para el conjunto de observaciones, se vuelve a calcular las distancias a los centroides de los clústeres y se reasignan a los que estén más próximos.

3. Se vuelven a recalcular los centroides de los k clústeres después de las reasignaciones de los elementos. Esto se obtiene mediante la realización de la media aritmetica para cada variable de su valor para los distintos elementos del clúster para encontrar así el punto central.

4. Se repiten los dos pasos anteriores hasta que no se produzca ninguna reasignación, es decir, hasta que los elementos se estabilicen en algún grupo o en su defecto se repiten por un número máximo de iteraciones determinado a priori.

Usualmente, se especifican k centroides iniciales y se procede al paso (2) y, en la práctica, se observan la mayor parte de reasignaciones en las primeras iteraciones.

El método k-medias es óptimo en problemas en los que se pretende reducir la distancia de los elementos dentro del clúster. Además ofrece un tratamiento relativamente rápido de los datos.

Por otro lado, una de las principales desventajas de este método es su dependencia de los clústeres elegidos inicialmente. Para mitigar el impacto de esta dependencia se puede efectuar el algoritmo reiteradas veces con condiciones iniciales diferentes para así comprobar la estabilidad del resultado. Otra opción es utilizar otro análisis de clústeres como el análisis jerárquico para determinar los clústeres iniciales y aplicar K-Means a partir de esa distribución inicial.

6. Elección de las variables macroeconómicas y bolsas a estudiar

La elección del análisis de clústeres a aplicar depende en gran medida de las variables estudiadas y la clase de análisis que se pretenda realizar.

Para estudiar las empresas se requiere un análisis de sus datos históricos a partir de los cuales sacar conclusiones. Una gran cuestión en el análisis de datos históricos es que intervalo temporal se utiliza para seleccionar estos datos. Este intervalo debe ser lo bastante amplio para ser representativo, pero no demasiado grande como para estar cogiendo datos tan antiguos que no tengan relación con la realidad.

El criterio a seguir es hacer coincidir el horizonte de los datos históricos con el horizonte temporal de la inversión. Por lo tanto, para fijar el histórico de datos a utilizar primero se debe determinar para qué tipología de inversión se realiza el estudio.

En general se suele clasificar el horizonte temporal en tres categorías diferentes:

-**Corto plazo**: Se consideran corto plazo inversiones con un horizonte temporal de hasta tres años. Estas inversiones a corto plazo suelen enfocarse con visión conservadora puesto que, al plantearse cerrar las posiciones en poco tiempo, no daría tiempo de recuperar la cartera en caso de que se produjesen fuertes movimientos iniciales a la baja.

-**Medio plazo**: Se consideran medio plazo inversiones con un horizonte temporal de entre 3 y 10 años. Estas inversiones se plantean balanceando el riesgo y la rentabilidad, dando mayor peso a la rentabilidad cuanto mayor sea el tiempo.

-**Largo plazo**: Se consideran largo plazo inversiones con un horizonte temporal de más de 10 años. Estas inversiones se plantean fijándose principalmente en la rentabilidad puesto que a pesar de que se produzcan movimientos de la cartera a la baja, hay tiempo para que los valores alcancen el valor deseado y margen de tiempo para cerrar posiciones en el momento que parezca más adecuado, sin depender de restricciones temporales.

Estas consideraciones generales siempre se han de tener en cuenta bajo el contexto de la tolerancia al riesgo del inversor.

Este trabajo analiza una inversión de la cual se reciben sus beneficios en un plazo de tiempo cercano, es decir, una inversión a corto plazo. Dentro del corto plazo se marca el horizonte temporal de tres años puesto que es un período temporal que permite asumir un cierto riesgo sin comprometer los beneficios de la cartera. En caso de tratarse horizontes temporales inferiores se entraría en riesgo de que un mal momento situacional causase pérdidas en la cartera. Por lo que podría hasta no ser recomendable invertir en empresas y ser más recomendable analizar otros mercados como la renta fija.

En este análisis, sobre un conjunto de empresas se estudiarán una serie de características que aporten información determinante sobre el desarrollo que han tenido estos negocios a lo largo de los últimos tres años des del punto de vista de los inver-

sores y los mercados financieros. Estos conjuntos de características se someterán a un análisis de clustering con la finalidad de definir conjuntos de empresas similares.

Con la finalidad de comparar empresas a través de las diferentes bolsas más relevantes a nivel mundial, se han elegido tres en concreto. La primera de ellas es la española por su interés como bolsa local. Las otras dos de ellas son representativas de los dos continentes con mayor peso en las finanzas:

-**IBEX 35**

-**Eurostoxx 50**

-**Dow Jones (Dow 30)**

En total, para realizar el estudio se han elegido de 115 empresas de los índices indicados.

Para analizar el parecido de estas empresas se compararan una serie de variables que están relacionadas con el interés que puedan tener las empresas de cara a los inversores.

Posibles características de los activos financieros seleccionables para estudiar son:

-**Rendimiento**: Es la principal característica de interés para un inversor y la que resume mejor el desarrollo de una empresa en los mercados financieros. El rendimiento es el beneficio obtenido por unidad monetaria invertida. Esta medida sirve para comparar las inversiones posibles entre diferentes opciones puesto que la finalidad última de invertir es obtener un beneficio de ello.

En última instancia es el rendimiento lo que determinará si una empresa es viable des de el punto de vista de los accionistas o no lo es.

Viene definida con la fórmula:

$$R = \frac{p_f - p_i}{p_i}$$

donde:

-p_f significa el precio final.

-p_i significa el precio inicial.

Se considera el horizonte temporal tal y como lo hemos definido anteriormente de tres años, haciéndolo coincidiendo además con los datos fiscales presentados por las empresas. Por lo que se elegirán datos de las fechas correspondientes a los años fiscales comprendidos en el intervalo entre 2015 y 2017.

Sabemos que estos datos no coincidirán con los años fiscales de todos los países analizados. Como por ejemplo Estados Unidos, cuyo año comprende de octubre a septiembre. Pero sí que se conoce que se producirán todos los eventos fiscales igualmente para ellos a pesar de estar desordenados.

Para la selección del rendimiento en lugar de calcular la diferecia entre precio inicial y final del período se calcularía semanalmente y después se realizaría la media de esas mediciones. Se aplicaría este críterio porque se considera que es el intervalo

sobre el cual una inversor con visión a tres años iria revisando su cartera. Realizaria esas inspecciones períodicas en medio de su período de inversion para así variar la composición de en caso de producirse serias modificaciones en el mercado financiero.

Por lo tanto, se considera que el dato de interés para el inversor no es tanto el rendimiento global como los rendimientos que observaria en cada inspección.

-**Volatilidad**: junto con el rendimiento es una de las variables más interesantes a la hora de analizarlas. Esta variable es un claro indicador del riesgo que se corre al invertir en una determinada empresa y lo acentuados que serán los movimientos de sus acciones.

En este sentido, activos con gran volatilidad pueden verse involucrados en variaciones de valor que supongan grandes diferencias tanto positiva como negativamente respecto al valor inicial.

Dado el estrecho vinculo entre el rendimiento y la volatilidad se ha usado conjuntamente con el rendimiento en multitud de modelos y teorías, un ejemplo de ellas es el CAPM (Capital Asset Pricing Model), donde se indica que en una inversión optima, un mayor rendimiento se consigue únicamente aceptado un mayor riesgo, es decir, con mayor volatilidad.

La volatilidad se medirá en términos de la desviación estándar que viene dada por la fórmula:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2}$$

Donde cada x_i es el valor de la empresa al cierre de cada sesión y x es la media.

-**Dividendos**: los dividendos son un valor muy representativo sobre la rentabilidad de la empresa. Tradicionalmente los dividendos se asocian al reparto de beneficios entre los accionistas. Es por ello que el hecho que una empresa pague un dividendo estable y de una cuantía considerable se tienda a relacionar con que dicho negocio marcha bien.

Puesto que las empresas conocen este hecho tienden a disociar los dividendos de los beneficios y a mantenerlos constantes y crecientes. Aun así, una de las principales causas de que un dividendo crezca de forma sostenida es la buena marcha del negocio.

Además, muchos economistas defienden la influencia del dividendo en el precio de las acciones. Su principal efecto sobre el precio es debido a que son uno de los métodos mediante el cual los accionistas recuperan su inversión. Por lo tanto, se puede concebir el rendimiento de una empresa como un bono perpetuo cuyos flujos de caja son los dividendos.

Se han creado diversos modelos con la intención de reflejar este hecho como el modelo de Gordon-Shapiro.

Los dividendos se pueden estudiar des de muchas perspectivas distintas, entre otros, se pude analizar des de la perspectiva de el dividendo como valor absoluto,

des de su valor relativo al precio de la acción dando lugar a la rentabilidad de la inversión si se hiciese para cobrar dividendos o des de la perspectiva de la proporción de dividendo que se reparte respecto al beneficio que ha obtenido la empresa aquel año, dando información sobre sus políticas financieras.

Se ha seleccionado analizar la proporción de los beneficios que la empresa reparte como dividendo, o payout ratio. Ya que para una inversión a corto plazo, el beneficio obtenido directamente por los dividendos que se cobran no es tan importante como las políticas de la empresa y cómo influyen estas a la evolución de precio de la acción.

El payout ratio se calcula de la siguiente manera:

$$payout\ ratio = \frac{dividendos\ por\ acción}{beneficios\ por\ acción} = \frac{dividenos\ totales}{beneficio\ neto}$$

-**Valor en libros de la empresa (book to market ratio)**: El book to market ratio da una medida objetiva sobre el valor de la empresa. Esta indica el dinero que recuperarían los inversores en caso de que la empresa cerrase y vendiese sus activos para pagar sus deudas.

Los detractores de este método como Khan and Zuberi [4] argumentan que es un método poco fiable porque la contabilidad es muy manipulable y está basada en conceptos abstractos como la depreciación.

Hay que tener en cuenta además que el índice puede tener una aplicación limitada en ciertas empresas contemporáneas. Ya que, en muchas de ellas, sus activos de más valor no son tangibles sino los que corresponden al personal y al conocimiento. Estas características que no aparecen en los libros de contabilidad están muy relacionadas con el desarrollo de la compañía.

Por otro lado, investigadores como Damodaran (2012) [3] apoyan el uso del método aportando estudios que relacionan un bajo índice con una elevada rentabilidad de las empresas.

-**Beneficio neto (Price Earning ratio)**: es un indicador sobre los resultados que ha obtenido la empresa y sobre su actividad a lo largo del año. A su vez permite predecir en qué medida esta podrá seguir dando rentabilidad a sus inversores.

El beneficio de la empresa se suele considerar una indicación relevante sobre el retorno obtenido de una empresa ya que este se puede destinar a dos usos.

-El primero de ellos son los dividendos: Usualmente una proporción de este beneficio se reparte en forma de dividendos. Cuando el inversor recibe dividendos de una acción está recuperando su inversión sobre la misma, además tienen la ventaja de que una vez obtenidos no se pueden perder, a diferencia de los incrementos en el precio de la acción que pueden sufrir variaciones.

-El segundo uso es la inversión: La inversión en nuevos activos o en pagar deudas implica que en el futuro la compañía será capaz de obtener beneficios aún mayores, lo que se traduce en un incremento de los dividendos. Este incremento esperado de los dividendos es una presión alcista para el precio de las acciones. En caso de que

los inversores vendiesen su parte, este caso les aportaría mayores beneficios.

El uso de la ratio en referencia al precio de la acción permite hacer comparables los ingresos de diferentes empresas de diferente tamaño. Ya que el tamaño de la empresa viene definido tradicionalmente por el precio de sus acciones y al dividir entre el mismo, obtenemos una medida para comparar los ingresos que obtendría un inversor en relación al capital que aportase.

Se calcula según:

$$PER = \frac{precio}{beneficio\ por\ acción}$$

-**Personal ocupado** El personal ocupado es un indicador del tamaño de la empresa. En numerosas clasificaciones se utiliza como el principal criterio para definir si la empresa es pequeña mediana o grande.

Se espera que empresas con un tamaño similar respondan de forma similar ante los movimientos macroeconómicos del mercado.

-**Volumen ventas** El volumen de ventas se encuentra relacionado directamente con los beneficios de la empras, además de estar relacionado con el tamaño de la empresa. De esta manera el volumen de ventas podría contarse como un factor determinante a la hora de clasificar empresas.

-**Máximo incremento** Uno de los criterios de rentabilidad para un inversor es la máxima cantidad de dinero que se podría obtener de una inversión. En este sentido, el mayor incremento en el precio de las acciones a lo largo del año indicaría cual habría sido la ganancia que se hubiese obtenido en caso de haber podido predecir el mercado correctamente y se hubiese podido entrar y salir de la inversión en los momentos adecuados.

-**Sharpe ratio** Es una medida de retorno en función del riesgo. El ratio contiene la diferencia entre el retorno esperado de una inversión y el retorno de un activo sin riesgo. Esta diferencia se divide entre la volatilidad, indicando cuanto beneficio adicional se obtiene por cada unidad de volatilidad o riesgo incurridos. Responde a la siguiente ecuación

$$Sharpe\ Ratio = \frac{r_p - r_f}{\sigma_p}$$

donde

-r_p significa el valor esperado del activo.

-r_f significa el retorno de un activo sin riesgo (risk free rate).

-σ_p significa la desviación estándar del activo.

Entre otras funciones indica si el exceso de beneficio esperado es debido a tomar mayores riesgos o es debido a mejores decisiones de inversión.

-**Value at risk**: otro indicador del riesgo que corre una empresa. Define la perdida que se asume en casos extremos de circunstancias adversas. Es usada la perdida

relacionada con el percentil 5 %.

Finalmente se ha optado por estudiar las variables de rentabilidad, volatilidad, dividendos, valor en libros de la empresa (book to market ratio) y beneficio neto (Price Earning ratio).

Las variables dividendos (payout ratio), valor en libros de la empresa (book to market ratio) y beneficio neto (Price Earning ratio) se han seleccionado por ser los factores que incitan a que una inversión sea rentable. Se ha seleccionado la rentabilidad adicionalmente, aun sabiendo que puede producirse una doble representación de este concepto en forma de altas correlaciones entre variables. Esta doble representación es en parte intencionada puesto se pretende que la clasificación represente principalmente este último valor.

Adicionalmente se valorará la volatilidad porque tal como se ha indicado anteriormente es otra de las variables fundamentales en cuanto a la clasificación de carteras.

7. Selección de la ponderaciones

De cara a la realización del análisis sobre las empresas, nos encontramos un problema a la hora de definir el método de análisis de clústeres a utilizar.

Las distintas variables definidas para el análisis son fácilmente calculables, pero se desconoce en que medida estas definen las características de la empresa subyacente. Este tema se combina con el hecho de que las variables están en escalas diferentes lo cual también es un factor a considerar

La gran mayoría de los métodos de clústering que emplean variables con una importancia diferente o de distinto tamaño, aplican una función de ponderación a las distintas variables con la finalidad de hacerlas comparables.

Muchos de esto métodos basan su análisis en las distancias entre elemento. En el caso de usar la distancia euclidea ponderada como distancia base, sean $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$ dos elementos a clasificar, la función distancia tiene el siguiente aspecto:

$$D_w(x, y) = \sqrt{\sum_{k=0}^{n} w_k(x_k - y_k)^2}$$

donde $w = (w_1, ..., w_n)$ indica la ponderación de las variables.

En estos casos esta función es elegida de forma arbitraria por quien realiza el análisis. Como se ha comentado anteriormente la funciones pueden ser variadas y comprende un amplio rango desde seleccionar pesos arbitrarios a hacer que la contribución media de cada variable al resultado sea la misma. Si bien en algunos casos se ha logrado ofrecer una solución al problema de que las variables tengan distintos tamaños, siguen existiendo grandes dificultades para valorar el aporte que deben tener las diferentes variables cuando tratan medidas con importancias diferentes.

Por ejemplo, en este caso de realizar clústeres sobre empresas que sus datos contuvieran entre otros el payout ratio y el book-to-market ratio, no está definido un método para determinar si la payout ratio es un indicador más apropiado de similitud, si lo es el book to market o si ambos tienen una importancia exactamente igual a la hora de comparar individuos.

Pudiendo resultar esto en una variedad de ponderaciones diferentes como por ejemplo seria que el payout debiera tener el doble de importancia en la distancia que el book to market, o a la inversa. En caso de que se calculase la distancia sin ponderaciones, se estaría considerando que ambas variables tienen exactamente la misma importancia a la hora de comparar empresas, lo cual podría o no podría ser cierto.

La principal implicación de esto es que los resultados sean altamente dependientes del criterio subjetivo del analista para decantarse por una ponderación u otra.

Con la finalidad de tratar este problema se ha desarrollado el método de la

partición más estable. Veamos como se ha definido este método.

8. Lemas necesarios

Denominamos X al conjunto finito de N elementos para organizar en clústeres. Cada elemento $x \in X$ presenta k variables que se representan $x = (x_1, .., x_k)$

Definición 8.1. *Denominaremos ponderación a un vector $\sigma \in R^k$, tal que $\sigma = (\sigma_1, ..., \sigma_k), \sigma_i \geq 0, \forall i$*

Definición 8.2. *Sea w una ponderación denominamos distancia euclidea ponderada por esa ponderación a una aplicación de la forma*

$$d_w(\ ,\ ) : X \times X \to \mathbb{R}^+$$

$$(x, y) \to \sqrt{\sum_{(i=1)}^{K} \sigma_i (x_i - y_i)^2}$$

Denominaremos D_Ω al conjunto de distancias ponderadas. Dada la forma de la definición es trivial observar que existe una biyección entre ponderaciones y distancias ponderadas. Esta será denominada

$$\phi : W \leftrightarrow D_\Omega$$

Definición 8.3. *Definimos una relación de equivalencia $\approx$ de las distancias. Dos distancias ponderadas d_{σ_1} i d_{σ_2} son $\approx$ equivalentes si :*

$$\forall x, y \in X, d_{\sigma_1}(x, y) = q d_{\sigma_2}(x, y),\ q \in \mathbb{R}^+ - \{0\},$$

Que cumplan esta propiedad es equivalente a decir que sus vectores de ponderaciones sean proporcionales. Veamos este hecho primero, es decir veamos que:

$$d_{:\sigma_1} \approx d_{\sigma_2} \Leftrightarrow \forall x, y \in X, d_{\sigma_1}(x, y) = q d_{\sigma_2}(x, y),\ q \in \mathbb{R}^+ - \{0\} \Leftrightarrow \sigma_1 = q^2 \sigma_2$$

Es cierto puesto que:

$$\forall x, y \in X, d_{:\sigma_1} = q d_{\sigma_2} \Leftrightarrow \sum_{i=0}^{n} \sqrt{\sigma_{1i}(x_i - y_i)^2} = q \sum_{i=0}^{n} \sqrt{\sigma_{2i}(x_i - y_i)^2} \Leftrightarrow$$

$$\Leftrightarrow \sum_{i=0}^{n} \sqrt{\sigma_{1i}(x_i - y_i)^2} = \sum_{i=0}^{n} \sqrt{q^2 \sigma_{2i}(x_i - y_i)^2} \Leftrightarrow \sigma_{1i} = q^2 \sigma_{2i}\ \forall i$$

Con probar que es relación de equivalencia para uno de los casos, habremos demostrado que lo es tanto para distancias como para ponderaciones:
-Reflexiva:

$$\forall \sigma \in,\ \sigma = 1\sigma$$

-Simetrica

$$\forall \sigma_1, \sigma_2 \in \Omega \; tq \; \sigma_1 \approx \sigma_2 \Rightarrow \exists q \in \mathbb{R}^+ - \{0\}, \; tq \; \sigma_1 = q\sigma_2 \Rightarrow \frac{1}{q}\sigma_1 = \sigma_2 \Rightarrow \sigma_2 \approx \sigma_1$$

-Transitiva

$$\forall \sigma_1, \sigma_2, \sigma_3 \in \Omega \; tq \; \sigma_1 \approx \sigma_2, \sigma_2 \approx \sigma_3 \Rightarrow \exists k, q \in \mathbb{R}^+ - \{0\}, \; tq \; \sigma_1 = q\sigma_2 \; \sigma_2 = k\sigma_3$$

$$\Rightarrow \sigma_1 = qk\sigma_3 \Rightarrow \sigma_1 \approx \sigma_3$$

Inducimos que es también una relación de equivalencia de las ponderaciones tal que dos ponderaciones σ_1 i σ_2 son $\approx$ equivalentes si :

$$\forall x, y \in X, \sigma_1 = q\sigma_2, \; q \in \mathbb{R}^+ - \{0\},$$

. **Classes de equivalencia**
Puesto que es una relación de equivalencia podemos formar clases de equivalencia en el conjunto de las ponderaciones $\sigma / \approx = [\sigma] = w$ elegiremos como representante de la clase el unico elemento de la forma $\sigma = (\sigma_1, ..., \sigma_k), \; tal \; que \; \forall i \in \{1, ..., k\} \; \sigma_i \leq 1$ i además $\exists i, \; \sigma_i = 1$
Veamos que el elemento existe y es único:

$$\forall \sigma, \exists \sigma' \; tq \; \sigma` = max\{\sigma_1, \ldots, \sigma_n\}$$

Calculamos: $\sigma = (\frac{\sigma_1}{\sigma`}, ..., \frac{\sigma_k}{\sigma`})$ de manera que el elemento σ que pertenece a la misma clase que w cumple las condiciones para ser el representante porque por un lado $\sigma_i = \frac{\sigma_i}{\sigma`} \leq 1$ y por el otro $\sigma_j = \frac{\sigma_j}{\sigma`} = 1$ Es único por el método de calcularlo. Dada una ponderación σ, existe un único$\sigma` = max\{\sigma_1, \ldots, \sigma_n\}$ . Hemos observado que $\frac{1}{\sigma`}\sigma$ da un representante tal y como hemos descrito.

Si observamos cualquier otra ponderación

$$\sigma_2 \in [\sigma], \sigma_2 \neq \frac{1}{\sigma`}\sigma \Rightarrow q \in \mathbb{R}^+ - \{0\}, q \neq \sigma` \; t.q. \; \sigma_2 = \frac{1}{q}\sigma$$

Hay dos posiblidades

$$q < \sigma` \Rightarrow i \; t.q. \; \sigma_{2i} = \frac{1}{q}\sigma_i = \frac{1}{q}\sigma` > 1$$

Que contradice la definición que habíamos dado del representante por tener valores superiores a 1

$$q > \sigma` \Rightarrow \forall \sigma_{2i}, \sigma_{2i} = \frac{1}{q}\sigma_i < \frac{1}{q}\sigma` < 1$$

Que contradice la definición que habíamos dado del elemento por no tener ningún valor igual a 1.

Por lo tanto el representante de clase así descrito existe y es único

Inducimos que se pueden formar clases de equivalencia en $D_\Omega/\approx$, siendo el representante de clase el único elemento d_σ *tal que* $\sigma = (\sigma_1, ..., \sigma_k)$, *tal que* $\forall i \in \{1, ..., k\}\ \sigma_i \leq 1$ i además $\exists i,\ \sigma_i = 1$

Entre otras propiedades, podemos observar que las clases son isomorfas entre ellas i que dentro de una misma clase, todos sus elementos mantienen la distancia relativa entre los elementos a los que se aplica. Siendo que para cualquieras dos distancias $d_{\sigma_1}, d_{\sigma_2} \in D_W, d_{\sigma_1} \approx d_{\sigma_2}$:

$$\forall x, y, z \in W,\ si\ d_{\sigma_1}(x, y) \leq d_{\sigma_1}(x, z) \Leftrightarrow\ d_{\sigma_2}(x, y) \leq d_{\sigma_2}(x, z)$$

Sea $W = D_w/\approx$ el conjunto de clases, W es isomorfo a

$$\cup_{i=1}^{n}[0, 1] \times .. \times [1]_i \times .. \times [0, 1]$$

Lema 8.4. *Sea X un conjunto, W un conjunto de clases de ponderaciones y sea:*

$$\phi : W \to \mathcal{P}$$
$$w \to P_w$$

donde P_w es la partición del conjunto X resultante de aplicar un método de análisis de clústeres cuyo criterio de asignación sea asignar cada elemento al clúster que más cercano, mediante el uso de la distancia d_w.

Sea w_1 y w_2 dos clases de ponderaciones tales que $\phi(w_1) = P_{w_1} = \phi(w_2)$ entonces para toda ponderación intermedia de la forma $w_\lambda = w_1 + \lambda(w_2 - w_1),\ 0 \geq \lambda \geq 1$ obtenemos $\phi(w_\lambda) = P_{w_1}$

Demostración Supongamos que existe una λ tal que $w_\lambda = P_\lambda \neq P_{w_1}$, llegaremos a contradicción.
Dado que

$$P_{w_\lambda} \neq P_{w_1} \Rightarrow \exists x, y, x \in X\ tq\ \begin{array}{r} d_{w_1}(x, y) < d_{w_1}(x, z) \\ d_{w_\lambda}(x, y) > d_{w_\lambda}(x, z) \\ d_{w_2}(x, y) < d_{w_2}(x, z) \end{array} \Rightarrow$$

$$\Rightarrow \begin{array}{r} \sqrt{\sum_{i=1}^{n} w_{1i}(x_i - y_i)^2} < \sqrt{\sum_{i=1}^{n} w_{1i}(x_i - z_i)^2} \\ \sqrt{\sum_{i=1}^{n} w_{\lambda i}(x_i - y_i)^2} > \sqrt{\sum_{i=1}^{n} w_{\lambda i}(x_i - z_i)^2} \\ \sqrt{\sum_{i=1}^{n} w_{2i}(x_i - y_i)^2} < \sqrt{\sum_{i=1}^{n} w_{2i}(x_i - z_i)^2} \end{array} \Rightarrow$$

$$C = \sum_{i=1}^{n} w_{1i}(x_i - y_i)^2 - \sum_{i=1}^{n} w_{1i}(x_i - z_i)^2 < 0$$

$$\Rightarrow \begin{array}{r} C = \sum_{i=1}^{n} w_{2i}(x_i - y_i)^2 - \sum_{i=1}^{n} w_{2i}(x_i - z_i)^2 > \\ > \sum_{i=1}^{n} \lambda(w_{2i} - w_{1i})(x_i - z_i)^2 - \sum_{i=1}^{n} \lambda(w_{2i} - w_{1i})(x_i - y_i)^2 \end{array} \Rightarrow$$

$$\begin{array}{r} C = \sum_{i=1}^{n} w_{2i}(x_i - y_i)^2 - \sum_{i=1}^{n} w_{2i}(x_i - z_i)^2 < \\ < \sum_{i=1}^{n}(w_{2i} - w_{1i})(x_i - z_i)^2 - \sum_{i=1}^{n}(w_{2i} - w_{1i})(x_i - y_i)^2 \end{array}$$

$$C < 0$$
$$\Rightarrow \quad C > \lambda(w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2)$$
$$C < (w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2)$$

Veamos que el caso $C > \lambda(w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2)$ contradice los otros dos. En este punto $(w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2)$ puede ser negativo, positivo o 0. En los casos positivo y 0 tenemos:

$$(w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2) \geq 0 \ y \ \lambda \geq 0 \Rightarrow$$

$$\Rightarrow C < 0 \leq \lambda(w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2)$$

Es una contradicción puesto que $C > \lambda(w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2)$. En el caso negativo tenemos:

$$C < (w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2) \leq 0 \ y \ 0 \leq \lambda \leq 1 \Rightarrow$$

$$\Rightarrow C < \lambda(w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2) \leq 0$$

, que se contradice con $C > \lambda(w_{2i} - w_{1i}) \sum_{i=1}^{n}((x_i - z_i)^2 - (x_i - y_i)^2)$.

Todos los casos terminan en contradicción por lo que debe ser $\Phi(w_\lambda) = P_{w_1}$, como queríamos demostrar.

Corolario 8.5. *Sea P una partición, $w_1, ..., w_l \in W$ un conjunto de clases de ponderaciones para N variables tal que $\Phi(w_i) = P \ \forall 1 \leq i \leq l$ y $A \subset W$ el polítopo de mayor volumen cuyos vértices forman parte de $w_1, ..., w_l$. Entonces $\forall \sigma \in A, \ \Phi(\sigma) = P$.*

Demostración En primer lugar se debe recordar que un polítopo es la generalización a cualquier dimensión de un polígono. Se define polítopo de dimensión d a la envolvente convexa de un subconjunto finito S del espacio euclideo $\mathbb{R}^d$. Sean $w_1, ..., w_s, \ s < l$ los vértices del polítopo, aplicando el lema 8.4 podemos observar que las ponderaciones que forman partes de las aristas cumplen $\Phi(w_i) = P$. Tras otra aplicación del lema, vemos que es cierto para las variedades de dimensión 2 con vértices en $w_1, ..., w_s$, Sucesivas aplicaciones del lema 8.4 nos permiten comprobar que lo mismo sucede para las variedades de dimensión 3, 4 y sucesivas. Hasta que en no más de N iteraciones se comprueba que la propiedad es cierta para todos los punto del polítopo.

9. Método de la partición más estable

El método de la partición más estable, o por sus siglas PMS, está diseñado específicamente para las situaciones en las cuales no está definido el peso relativo de las variables. Además permite aportar un criterio diferente a la hora de establecer ponderaciones, criterio que no se basaría en el conocimiento o intuiciones subjetivas del analista.

Para hacer esto, el método de la partición más estable pretende invertir la tradicional secuencia de pasos de en primer lugar definir una ponderación y a continuación elaborar un análisis de clústering. La secuencia según este método pasaría a ser encontrar el clúster deseado entre todos los posibles y a continuación, en caso de que fuese necesidad del analista, podría definir cuáles son las ponderaciones que dan lugar a dicho clúster.

El criterio para seleccionar clústeres en este caso es el siguiente: Una partición se considerará la mejor clasificación de entre las posibles cuando sea la partición resultante de un mayor número de ponderaciones. Una vez definida la partición se podrá concretar que ponderaciones le han dado lugar.

Se efectuaran los siguientes pasos:

1-Definir un método de clústering. Este método de clústering será con el que se creara una distribución en clústeres de las variables con cada una de las distintas distancias. Estas distancias serán las resultantes de modificar la distancia usada por el método según marquen las ponderaciones.

2-Computar el método para las distintas ponderaciones existentes. Habitualmente, debido a que el vector de ponderaciones se moverá sobre variables continuas, resultará imposible computar el método para absolutamente todas las ponderaciones ya que serán infinitas. En su defecto se efectúa el análisis para una cantidad suficiente de ponderaciones que nos permita obtener conclusiones sobre la distribución de las particiones en el espacio de ponderaciones.

3-Calcular el volumen del espacio de ponderaciones que da lugar a cada partición.

Este se puede deducir a partir de localizar todas las ponderaciones que aplicando el método de análisis con esas ponderaciones se produce la misma distribución de los elementos siguiendo los mismos clústeres.

Una vez se han localizado estos puntos en el espacio de ponderaciones, se puede reproducir el polítopo de mayor tamaño posible que los tiene como aristas. Y tal como se ha visto en el corolario 8.5, se puede asegurar que todos los puntos de su interior dan lugar corresponden a ponderaciones que dan lugar a la misma distribución de clústeres.

Finalmente, una vez definido el polítopo, se puede computar su volumen.

4-Seleccionar aquel clúster que tiene mayor volumen.

Por último, se selecciona como clúster resultante del método, aquel clúster que se ha encontrado que se da para un mayor volumen de puntos.

La principal dificultad del método es el esfuerzo computacional que supone calcular reiteradamente el mismo análisis en el punto 2 para suficientes ponderaciones como para poder obtener conclusiones.

Con el objetivo de solucionar ese problema se considera óptimo utilizar un método de análisis de clúster computacionalmente rápido.

Se ha propuesto un método de clústering con el objetivo de aportar esta simplicidad de cálculos, además de eliminar el impacto de los diferentes tamaños de las variables. Este método consiste en asignar cada elemento al clústeres del elemento más cercano. Además, con la finalidad de eliminar la influencia de los diferentes tamaños de las variables, esta distancia se calcula porcentualmente.

Pasos del análisis de clúster:

1-Seleccionar el primer elemento.

2-Calcular la distancia a los demás elementos Esta distancia se realizaría en forma de porcentaje, es decir, para cada variable se computaría cual es la diferencia porcentual frente a la misma variable del elemento a comparar. Es decir la distancia entre dos elementos para una única variable seria $d(x_j, x_i) = \frac{x_i}{x_j} - 1$ Esta distancia se calcularía de forma agregada y ponderada. Para un elemento x_i respecto a x_j la distancia seria $d_w(x_j, x_i) = \sum_{k=1}^{n} w_k(\frac{x_{i,k}}{x_{j,k}} - 1)$

3-Localizar el elemento más cercano y añadir este elemento más cercano al clúster del elemento estudiado. En caso de que el elemento más cercano estuviese ya en un clúster, se unirían ambos clústeres.

4-Seleccionar el siguiente elemento y repetir los pasos 2 y 3 hasta que se hayan seleccionado todos los elementos.

Las ventajas de este método serian tanto su aparente velocidad como el hecho de que las distancias no dependen de las unidades en que se mida cada distancia.

Otra alternativa planteada es el uso de K-Means. K-Means es conocido por ser uno de los algoritmos de clústering más rápidos a la hora de realizar las clasificaciones. Se plantea la posibilidad de que gane mayor velocidad a pesar de hacer un mayor número de iteraciones debido a que las comparaciones no se realizarían de un elemento frente a todos los demás sino de un elemento frente a un número reducido de centros. En el siguiente punto se procederá a ver en detalle cuál de los dos métodos de clústering complementa mejor al método de la partición más estable cuando se busca velocidad.

10. Número de operaciones

El número de operaciones a realizar en el método de la partición más estable está directamente determinado por el número de elementos de la muestra n, el número de variables v y la precisión del estudio P. Así como el algoritmo de clústering a aplicar.

Para decidir si aplicar el método de K-Means o si usar el algoritmo de agrupar cada elemento al clúster del elemento que tiene más cercano se mira el tiempo de cálculo de ambos métodos.

Agrupación al elemento más cercano

Por cada valor de ponderación se han de computar las distancias de los n elementos dando lugar a una matriz n*n simétrica con diagonal 0. Se requieren $\sum_{i=0}^{n} i$ cálculos de distancias. En cada uno de ellos se calcula la diferencia, se eleva al cuadrado y se suma, resultando en 3v-1 operaciones por distancia.

Por tanto por cada peso se necesitan $(3v-1)\sum_{i=0}^{n} i = (3k-1)\frac{(n+1)n}{2}$ cálculos. Se añaden (n-1)(n-2) comparaciones. Para n grandes, esta valor aumenta el número de operaciones en

$$(3v-1)\frac{(n+1)n}{2} + (n-1)(n-2) \approx (3v-1)\frac{(n+1)n}{2} + (n+1)n =$$

$$= (3v+1)\frac{(n+1)n}{2}$$

Hasta el momento los cálculos son del orden de $\frac{3}{2} * v * n^2$.

Esto se realiza para cada peso por lo que se efectúa $vP^{(v-1)}$.

El orden total tiende a $\frac{3}{2}v^2n^2P^{(v-1)}$.

Se observa que el número de operaciones es muy sensible al valor de v. Por lo que para evitar la imposibilidad operativa de calcularlo se necesita mantener el número de variables limitado. Esta sería una de las principales limitaciones del método.

K-Means

El número de operaciones a realizar en el método de la partición más estable esta directamente determinado por el número de elementos de la muestra n, el número de variables k y la precisión del estudio P y adicionalmente, está influido por el número de centros o clústeres K.

Para cada valor de ponderación se han de computar las distancias de los n elementos a cada uno de los C centros dando lugar a una matriz n*K. Se requieren $K * n$ cálculos. Cada uno de ellos se calcula la diferencia se eleva al cuadrado y se suma, resultando en 3v-1 operaciones por distancia.

Por tanto para cada peso se necesitan $(3v-1)*n*K$ cálculos. Se añaden (n-1)*K comparaciones. Hasta el momento los cálculos $(3v-1)*n*K+(n-1)*K3v*n*K$

K-Means es un método que converge rápidamente, habitualmente en 4 o 5 iteraciones por lo que se realizaría una aproximación de $5 * 3v * n * K = 15vnK$

operaciones por cada aplicación de K-Means.

Por lo tanto, por cada iteración los cálculos son del orden de $v * n * K$.

Esto se realiza para cada peso por lo que se efectúan $v * P^{(v-1)}$ operaciones. El orden total tiende a $15v^2nP^{(v-1)}K$.

Se observa que a pesar de utilizar K-Means dentro del método de la partición más estable, el método sigue siendo muy sensible al valor de v. Por lo que se continúa manteniendo la limitación sobre las variables analizadas.

Con respecto al método de clústering anterior hay que remarcar la reducción de dependencia respecto al número de elementos n. Siendo que el número de elementos de la muestra anteriormente influía sobre el número de operaciones de forma cuadrática y actualmente lo hace de forma lineal. Por el contrario en K-Means este número de operaciones también depende del número de centros K y la constante se ha incrementado de $\frac{3}{2}$ a 15.

Podemos observar que

$$15v^2nP^{(v-1)}K < \frac{3}{2}v^2n^2P^{(v-1)}K \Leftrightarrow 15nK < \frac{3}{2}n^2 \Leftrightarrow K < \frac{n}{10}$$

Por lo tanto, para valores de n elevados y de K reducidos, utilizar el método con K-means permite hacer menos operaciones.

En particular para este método se plantea el uso de aproximadamente 110 empresas y 5 clústeres por lo que se cumpliría este punto.

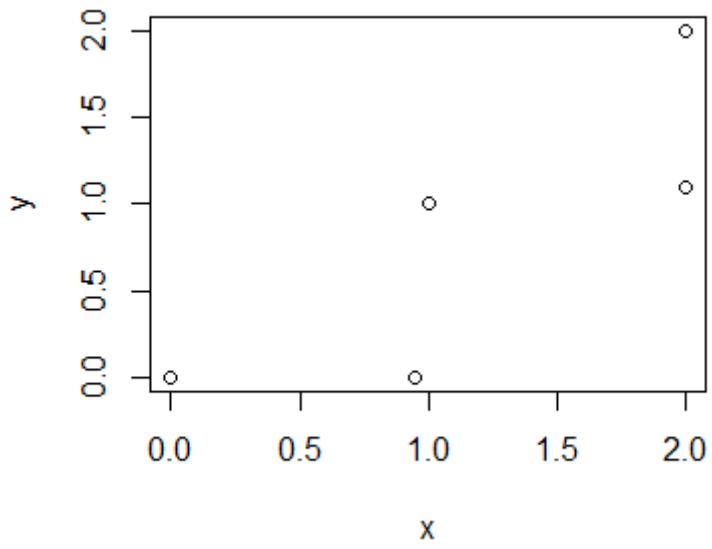


Figura 2: Distribución de los puntos

11. Comparativa

Tras observar que el método de la partición más estable obtendría resultados más rápidamente con K-Means, se procede a realizar una comparativa a nivel teórico e ilustrativo de cómo funcionaria el método y como mejoraría la predicción de K-Means.

Sea $X \in \mathbb{R}^2$ un conjunto de elementos a clasificar, que suponemos de la forma

$$X = \{a = (0,0), b = (0{,}95, 0), c = (1,1), d = (2,1{,}1), e = (2,2)\}$$

Este conjunto se puede representar tal y como se ve en la figura 2.
Aplicaremos los métodos de K-Means y partición más estable seleccionando como núcleos iniciales de los clústeres los puntos a, c y e.

Aplicando K-Means

En primer lugar hay que evaluar la distancia a los núcleos. Estas distancias se reflejaran en una matriz donde la columna será distancia al primer núcleo en este caso el (0,0), la segunda será la distancia a (1,1) y la tercera será la distancia a (2,2).

	*Distancia a c*1	*Distancia a c*2	*Distancia a c*3
a	0	$1,41$	$2,82$
b	$0,96$	$1,00$	$2,26$
c	$1,41$	0	$1,41$
d	$2,286$	1	$0,9$
e	$2,82$	$1,41$	0

El siguiente paso es asociar cada punto al clúster del centro más cercano. Se puede observar que en la primera iteración los puntos a y b están más cerca del primer centro mientras c y d formarían parte del segundo clúster.

Quedaría $a, b \in K_1$, por otro lado $c \in K_2$ y finalmente $d, e \in K_3$ El siguiente paso de K-Means consiste en asignar los nuevos centros a cada clúster. Siendo

$$c1 = \frac{a+b}{2} = (0{,}48, 0)$$

$$c2 = c = (1, 1)$$

$$c3 = \frac{d+e}{2} = (2, 1, 55)$$

A continuación, puesto que la iteración anterior ha sido la primera, se deberá repetir de nuevo la iteración y encontrar de nuevo cual es el centro más cercano de cada punto.

	Distancia a $c1$	*Distancia a* $c2$	*Distancia a* $c3$
a	$0,48$	$1,41$	$2,53$
b	$0,48$	$1,00$	$1,87$
c	$1,13$	0	$1,14$
d	$1,876$	1	$0,45$
e	$2,51$	$1,41$	$0,45$

Podemos ver por lo tanto que se vuelve a repetir la situación anterior en la cual a y b están más cerca de c_1 , c más cerca de c_2 y, d y e más cerca de c_3.

Se deberían volver a calcular los centros pero se puede deducir que si no han cambiado los clústeres, entonces los centros son los mismos.

Además una vez se ha repetido la distribución de los clústeres, se da el método por concluido y se observa que la distribución en clústeres resultante es $K_1 = a, b$, y $K_2 = c$ y $K_3 = d, e$.

Aplicación de la partición más estable

Aplicamos a continuación el método de la partición más estable para comparar el resultado.

En primer lugar hay que observar el espacio de las ponderaciones, que para dos variables es

$$W = \{1\} \times [0, 1] \cup [0, 1] \times \{1\}$$

A continuación se debe decidir cada que longitud se evaluaran los intervalos $\{1\} \times [0, 1]$ y $[0, 1] \times \{1\}$. En este caso se considerarán que subintervalos de longitud equivalente a 0,1 puntos básicos serán suficientes. Es decir se dividiera el primer intervalo en los siguientes puntos:

$$A = \{\{1\} \times \{0\}, \{1\} \times \{0, 1\}, \{1\} \times \{0, 2\}, \{1\} \times \{0, 3\}, \{1\} \times \{0, 4\}, \{1\} \times \{0, 5\},$$

$$\{1\} \times \{0, 6\}, \{1\} \times \{0, 7\}, \{1\} \times \{0, 8\}, \{1\} \times \{0, 9\}, \{1\} \times \{1\}\}$$

y el segundo de forma equivalente.

Se comienza a evaluar los intervalos fijando e igualando a 1 la componente w_x de la ponderación $w = (w_x, w_y)$ y iterando la segunda componente. La primera ponderación es $w = (1, 0)$, se aplica K-Means con esta ponderación de las distancias.

Recordamos que los centros iniciales son a=(0,0), c=(1,1) y e=(2,2). Se calculan de nuevo las distancias de cada punto hasta los centros.

	*Distancia a c*1	*Distancia a c*2	*Distancia a c*3
a	0	1	2
b	0,96	0,04	1,04
c	1	0	1
d	2	1	0
e	2	1	0

Se puede ver que tras la primera iteración se agrupa cada elemento con el centro más cercano, los clústeres quedarían como $k_1 = a, k_2 = b, c$ *y* $k_3 = c, d$

Se recalculan los centros que quedan como:

$$c1 = a = (0, 0)$$
$$c2 = \frac{b + c}{2} = (0{,}98, 0{,}5)$$
$$c3 = \frac{d + e}{2} = (2, 1{,}55)$$

Se recalculan las distancias:

	*Distancia a c*1	*Distancia a c*2	*Distancia a c*3
a	0	0,98	2
b	0,96	0,02	1,04
c	1	0,02	1
d	2	1,02	0
e	2	1,02	0

Se observa que se mantiene la misma distribución en clústeres y concluimos que el clúster asignando para la ponderación $w = (1, 0)$ es $K_{(1,0)} = \{\{a\}, \{b, c\}, \{d, e\}\}$ y lo nombraremos P_1.

A continuación se selecciona la siguiente ponderación, que es $w = (1, 0{,}1)$ Tras aplicar K-Means con esta ponderación se vuelve a obtener la distribución anterior y se concluye $K_{(1,0,1)} = \{\{a\}, \{b, c\}, \{d, e\}\} = P_1$.

Se repite el mismo procedimiento para el resto de ponderaciones tales que $w_x = 1$ y se obtienen los siguientes resultados, presentes en los cuadros 1 y 2.

Ponderación	(1,0)	(1,0.1)	(1,0.2)	(1,0.3)	(1,0.4)	(1,0.5)
Resultado	P_1	P_1	P_1	P_1	P_1	P_1

Cuadro 1: Particiones en el primer intervalo

Ponderación	(1,0.6)	(1,0.7)	(1,0.8)	(1,0.9)	(1,1)
Resultado	P_1	P_1	P_1	P_1	P_2

Cuadro 2: Particiones en el segundo intervalo

Donde P_2 es la partición de la forma $P_2 = \{\{a, b\}, \{c\}, \{d, e\}\}$.

Tras esto se efectúa el método de K-Means para las ponderaciones de la forma $w_y = 1$ obteniendo los clústeres de los cuadros 4 y 5.

Ponderación	(0,1)	(0.1,1)	(0.2,1)	(0.3,1)	(0.4,1)	(0.5,1)
Resultado	P_3	P_3	P_3	P_3	P_3	P_3

Cuadro 3: Particiones en el tercer intervalo

Ponderación	(0.6,1)	(0.7,1)	(0.8,1)	(0.9,1)	(1,1)
Resultado	P_3	P_3	P_3/P_2	P_2	-

Cuadro 4: Particiones en el cuarto intervalo

Donde P_3 es la partición de la forma $P_3 = \{\{a, b\}, \{c, d\}, \{e\}\}$. Y donde la ponderación (0.8,1) tiene un punto que equidista de dos centros y en función de una asignación o la otra, se obtienen dos particiones diferentes. Se trata pues del límite entre una partición y la siguiente.

Además, la ponderación (1,1) no se ha calculado nuevamente puesto que ya había sido calculada en la tabla anterior.

Con esto se puede ver que la distribución de las particiones es la que se muestra en el cuadro 5.

Partición	Apariciones
P_1	10
P_2	3
P_3	9

Cuadro 5: Distibución

La partición resultante de la aplicación del método de la partición más estable seria, por lo tanto, $P_1 = \{\{a\}, \{b, c\}, \{d, e\}\}$. Además se puede observar que el intervalo máximo que puede ocupar P_3, es $I(P_3) = [0, 0{,}8) \times \{1\}$. Mientras que el mínimo intervalo de P_1 es $I(P_3) = \{1\} \times [0, 0{,}9 + \epsilon)$. Por lo cual se confirma que en ninguna circunstancia P_3 seria más estable que P_1.

Es simple observar que $P_2 = \{\{a, b\}, \{c\}, \{d, e\}\}$, que era la partición resultante de K-Means tradicional, es una partición claramente menos estable que las anteriores. Para comprender el fenómeno con más detalle podemos observar que si $w = (w_x, w_y)$ es una ponderación, entonces, para $w_x < \delta = 0,9^2 - 0,1^2 = 0,8$. Entonces, cuando $w_y = 1$ tenemos:

$$d_w(c, d) = \sqrt{w_x(2-1)^2 + (1-1,1)^2)} < \sqrt{w_x(2-2)^2 + (2-1,1)^2)} = d(d, e)$$

Por otro lado

$$d_w(c,b) = \sqrt{w_x(1-0,96)^2 + (0-1)^2} > 1$$

$$1 > 0,96 = \sqrt{w_x(0-0,96)^2 + (0-0)^2} = d(a,b)$$

Por lo que los clústeres quedarían: $K_1 = \{a, b\}$, y $K_2 = \{c, d\}$ y $K_3 = \{e\}$.

Se puede observar que un fenómeno similar sucede cuando $w_y < \delta = 0,96^2 - 0,04^2 = 0,92$ y $w_y = 1$.

En cuyo caso los clústeres resultantes de aplicar K–Means con este conjunto de ponderaciones seria: $K_1 = \{a\}$, $K_2 = \{b, c, \}$ y $K_3 = \{d, e\}$ Una vez concluida la comparación teórica y viendo que el uso de la partición más estable puede darnos mejores resultados que K-Means, se procede a analizar el programa que realizara el análisis con este método.

12. Explicación programa

Para la ejecución del análisis se ha efectuado un programa en R. Una de las ventajas de esta herramienta de programación es su amplia aplicación en distintas sociedades e instituciones para el tratamiento de base de datos, el hecho de tratarse de código abierto, gratuito y con una creciente base de paquetes de información promovidos por los usuarios.

El principal inconveniente que presenta R es que es un lenguaje de ejecución lenta, este hecho se suple en muchos paquetes de funciones mediante la programación de ciertas partes en otros lenguajes que puedan procesar esas funciones más rápidamente.

El programa consta de diferentes partes: Lectura de datos, selección de ponderaciones, método de clústering, selección de la mejor ponderación.

-Lectura de datos: La lectura de los datos se ha efectuado a través de la función de R quantmod() [7]. Con esta función se puede seleccionar el ticker (que es el identificador de tres o cuatro caracteres asociado a cada empresa) de la empresa que se planea utilizar y descargar los datos de las cotizaciones de dicha empresa.

La función quantmod() dispone de campos para introducir que permiten seleccionar de que empresa se desea descargar sus datos, también se pueden descargar otra clase de información económica como podría ser la tasa de desempleo. Además de introducir la fuente de información que puede ser yahoo, google finance, la FRED, etc...

La misma función dispone de la opción de seleccionar un intervalo concreto de fechas para descargar los datos correspondientes a ese intervalo. En este caso se seleccionan los datos correspondientes al intervalo entre 01-01-2014 y 31-12-2017. Como portal de información se ha optado por yahoo finance.

Esta función se utiliza en combinación con la función saveSymbols que permite guardar en el ordenador los datos descargados anteriormente y poder tratar con ellos posteriormente.

Para los datos que no se han podido guardar utilizando las funciones anteriores se ha procedido a la descarga y la introducción en el programa manual de dichos datos. Se ha utilizado la misma fuente de información, que es yahoo finance, para obtenerlos.

Debido a la disponibilidad de los datos el análisis se ha efectuado finalmente sobre 93 empresas que se distribuyen de la siguiente forma: 34 empresas del Ibex, 30 empresas del Dow Jones, 29 empresas de Euros Stoxx.

Adicionalmente a la captación de datos, se ha realizado un análisis secundario para seleccionar el número de clústeres. Para ello se ha aplicado la función de R silhouette()

Esta función ofrece la posibilidad de obtener el coeficiente de silhouette de la partición resultante de aplicar un análisis de clústeres a un conjunto X. Este coeficiente es un indicador de la distancia entre clústeres, se obtiene realizando el

siguiente cálculo para cada punto:

$$Silhouette_p = \frac{B - A}{max(A, B)}$$

donde:

-A es la medida de la cohesión del clúster, definida como la media aritmética de la distancia entre el punto p y los demás puntos del clúster.

-B es una medida de la separación entre clústeres. En concreto es la media aritmética de la distancia que hay entre p y cada uno de los puntos del clúster más cercano.

Finalmente, para obtener el coeficiente global de la partición, se realiza la media de los coeficientes obtenidos para cada punto:

$$Silhouette = \sum_{p \in X} Silhouette_p$$

Por tanto cuanto mayor sea el valor obtenido en este coeficiente, mejor definido esta cada clúster. Esto acaba implicando que la partición en clústeres obtenida por el método aporta una mayor información sobre las variables.

Puesto que lo que se pretende obtener es un método que de información sobre que número de clústeres utilizar en el análisis a través de las distintas ponderaciones, no es coherente realizar la función silhouette para distintas ponderaciones. Se opta por seleccionar una ponderación para hacer el análisis silhouette. Esta ponderación es la que da lugar a la distancia euclidea, se selecciona tanto por coherencia con los métodos habituales de análisis así como porque todas las ponderaciones están realizadas a partir de intervalos que toman la distancia euclidea como base.

Por lo tanto el análisis se realiza aplicando K-Means a los datos en la distancia euclidea para diferentes números de clústeres. De cada aplicación de K-Means se obtiene un coeficiente de Silhouette.

La figura 3 es un ejemplo del resultado obtenido por la función silouette. En este caso al aplicarla al conjunto de las empresas aplicando K-Means con una distribución en 8 clústeres.

Tras diversas iteraciones se puede obtener un esquema de los resultados de silhouette en función del número de clústeres. Se muestra en la figura 4 con el número de clústeres en el eje horizontal y el coeficiente en el eje vertical sintetiza los resultados obtenidos.

El detalle de los resultados de aplicar silhouette para los distintos gráficos se encuentra en el anexo 2 parte 1. El detalle del programa para ejecutar silhouette se encuenta en el anexo 1 parte 3. De este grafico podemos observar que a medida que crece el número de clústeres, el coeficiente es creciente. Esta tendencia se mantiene hasta alcanzar el valor de 6 clústeres. A partir de ahí decrece irregularmente, sin superar nunca el valor del coeficiente para 6 clústeres.

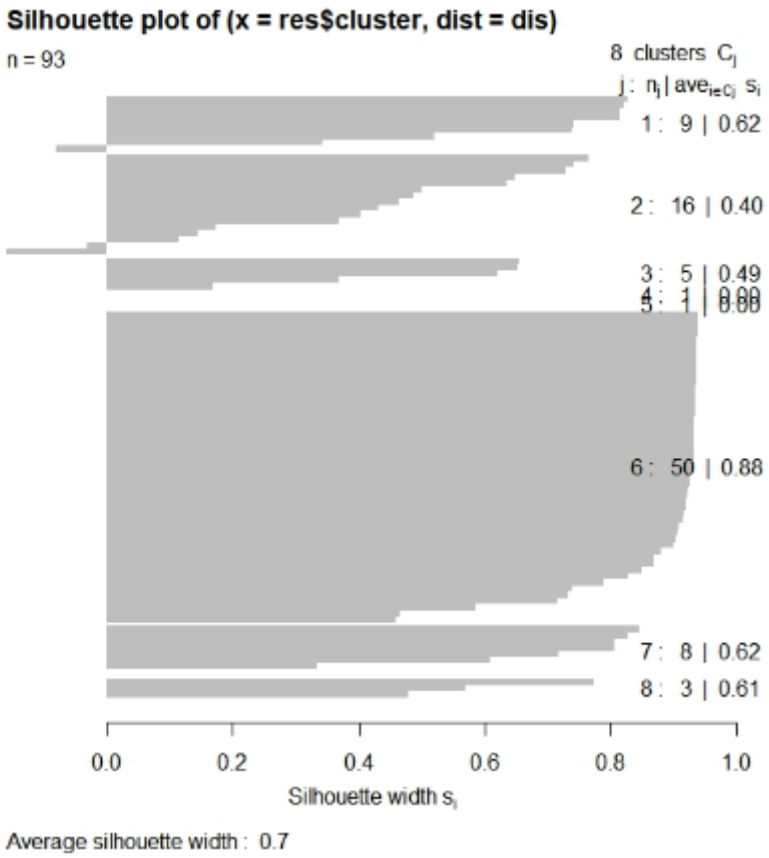


Figura 3: Resultados de silhouette para K=8

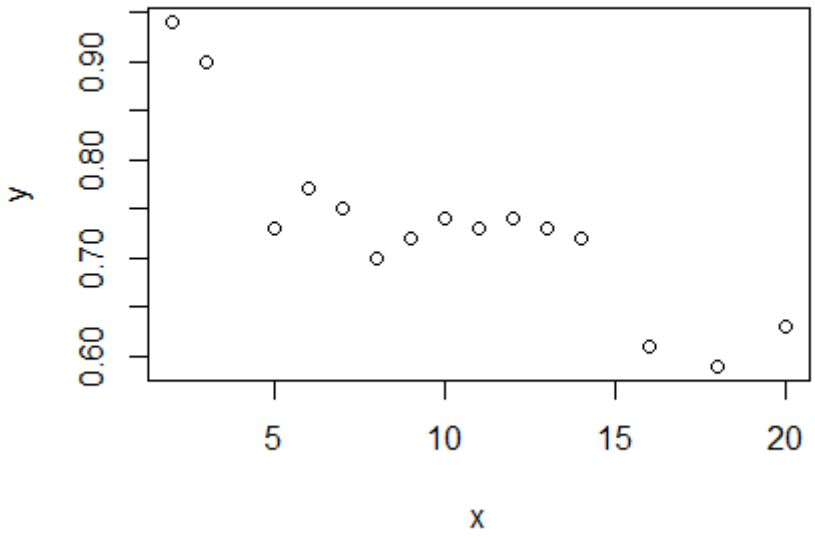


Figura 4: Evolución de los coeficientes de silhouette

Por lo tanto se selecciona 6 como el número de clústeres para el análisis por ser el valor máximo de silhoutte una vez descartados los clústeres iniciales. Para estos los valores elevados son un efecto de la distorsión creada por concentrar todos los elementos en un número excesivamente pequeño de clústeres.

-Método de clústering: En este segmento del programa se trata el método del análisis de las k medias. Lo aplica a los datos de la empresa captados en el primer apartado y en función de las ponderaciones.

R dispone de la función

$$kmeans(x,\ centers,\ iter.max,\ nstart,\ algorithm\ =$$

$$c("Hartigan-Wong",\ "Lloyd",\ "Forgy",\ "MacQueen"))$$

que realiza el análisis de clústeres aplicando dicho método, donde

-X es la matriz de datos numéricos sobre la que se realizará el análisis.

-Centers es el número de clúster con el cual se efectuará el análisis o en su lugar un número inicial de puntos distintos que servirán como centros de los clústeres. En caso de que se haya introducido únicamente un número, los centros de los clústeres se seleccionaran escogiendo k filas aleatorias de x.

-Iter.max es el máximo número de iteraciones que podrá efectuar el método

-Nstart es un valor designado únicamente para cuando "centers" es un número. Nstart indica el número de sets aleatorios de centros que serán evaluados para obtener el mejor resultado posible.

-Algorithm indica que variante dentro de K-Means se aplicará.

Esta función, entre otra información, devuelve un vector que indica a que clúster pertenece cada punto y la matriz de los centros de los clústeres.

Para la realización de este trabajo, esta función tiene la limitación que no permite incorporar ponderaciones en la distancia aplicada. Siempre utiliza la distancia euclidea equiponderada.

Antes este hecho se plantean dos posibles soluciones:

-Por un lado la realización de una función alternativa de K-Means que incorpore en sus cálculos las ponderaciones deseadas.

- La otra alternativa consiste en adaptar la función para que la matriz utilizada para el cálculo de K-Means incluya tanto los valores de las variables como las ponderaciones. Es decir, aplicar una función previa a K-Means para que los datos estén adaptados a las ponderaciones.

Ambas alternativas son viables así que se ha procedido a un análisis de las dos para determinar cual haría el programa más veloz.

Como análisis previo del problema se debe hacer notar que el programa 1 está diseñado parcialmente en R y parcialmente en otros lenguajes de programación para optimizar su tiempo de computación mientras el programa 2 está diseñado íntegramente en R. A priori el programa 1 debería ser más rápido y únicamente resultaría más lento en caso de que el ajuste externo para incluir las ponderaciones

fuese excesivamente lento computacionalmente.

Método 1: Función adicional

Para la incorporación de la función se ha optado por realizar cálculos anteriores de manera que las variables que llegasen a K-Means estén preparadas para que una vez se calculen distancias euclideas con ellas, estas distancias euclideas sean las equivalentes a una distancia euclidea ponderada por la ponderación w.

Se ha seguido el siguiente razonamiento:

$$D_w^2(x,y) = \sum_{i=1}^{n} w_i(x_i - y_i)^2 = \sum_{i=1}^{n}(\sqrt{(w_i)}x_i - \sqrt{(w_i)}y_i)^2 = D^2((\sqrt{(w)}x),(\sqrt{(w)}y))$$

Donde:

$$(wx) = (w_1 \times x_1, \ldots, w_n \times x_n)$$
$$(wy) = (w_1 \times y_1, \ldots, w_n \times y_n)$$

Por lo tanto el método consistirá en multiplicar cada variable por un factor correspondiente a su ponderación. Puesto que el factor depende únicamente de la variable elegida y no de la empresa correspondiente, se debe multiplicar cada columna por el mismo factor.

La forma más práctica de efectuar estas operaciones es multiplicar por la derecha una matriz diagonal con los factores correspondientes a cada variable.

En R esto sería:

```
1 x<-c(1,1,1,1,1)

3 kmeans(M\%*\%diag(sqrt(x)), 5, nstart = 1)
```

Método 2: Nueva función K-Means

Esta solución pasa por programar una función de K-Means completamente nueva, tal y como se muestra en el código del anexo 1 parte 4.

Se debe remarcar que funcionando conjuntamente con el resto del programa, este código captaría los datos enviados por el programa. Sin embargo para este diseño experimental se definen unas variables aleatorias que harán la función de datos.

Comparación de tiempos

Para la comparación de ambos métodos se ha procedido a realizar una prueba comparativa sobre el tiempo que tarda cada método en procesar un conjunto de datos aleatorios.

Para estos cálculos se utiliza la función de R "Sys.time()" que devuelve la hora a la que el sistema ha procesado ese punto del programa. Se utiliza una al inicio y otra al final del programa. Siendo de esta forma el tiempo empleado equivalente al tiempo entre las funciones.

Adicionalmente, para eliminar posibles desviaciones a la hora de elegir un set de datos, se utilizaran 10 iteraciones de cada función y se calculara la media.

Se fija el número de elementos a examinar en el test en 150 que es un número de elementos consistente y cercano al que se plantea el estudio del mercado. El número de variables de cada elemento es 5 puesto que son las que se analizarán.

Los resultados obtenidos son los siguientes:

Para el primer método son los presentes en el cuadro 6.

Iteración	tiempo (seg)
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
Media	0

Cuadro 6: Tiempos método 1

En cambio para el segundo método son los presentes en el cuadro 7.

Iteración	tiempo (seg)
1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0
9	0
10	0
Media	0

Cuadro 7: Tiempos método 2

Se puede observar que ambos códigos han tardado menos de 0 segundos en procesar los datos, esto se debe a que los programas son tan veloces procesando estos datos que no llegan a ser detectados por el umbral mínimo de la función "Sys.time()".

Habitualmente esto significaría que ambos códigos son igualmente validos para introducirlos en el programa pues la diferencia entre ambos es ínfima. Aun así el

código seleccionado se ejecutara prácticamente 50.000 veces, por lo que una diferencia de una decima de segundo podría llegar a significar una hora más de procesado.

En vista de la importancia de seleccionar el más veloz de ambos métodos se procede a un segundo testeo.

Este testeo consiste en ir elevando paulatinamente el número de elementos a los que aplicar K-Means con la finalidad de determinar la evolución de los tiempos de procesado, y deducir de ahí cual es el método que procesa los elementos de forma más veloz.

Para esto se ha iterado el método para 150 elementos y posteriormente el número de elementos se ha ido duplicando. Finalmente se ha realizado para 300, 600, 1200, 2400 y 4800 elementos.

El resumen de los resultados obtenidos se muestra en el cuadro 8, mientras que los detalles de los tiempos de cada ejecución se explicitan en el anexo 2 parte 3.

Elementos	Metodo 1	Método 2
150	0	0
300	0,1	0,5
600	0,2	0,5
1200	0,3	0,7
2400	0,5	1,3
4800	3	5,1

Cuadro 8: Comparativa tiempos de ejecución

De estos resultados se puede ver que aunque al ejecutar el programa para 150 elementos ambos métodos mostraban un comportamiento similar, a medida que se incremente el número de elementos, el método de adaptar los datos anteriormente, se muestra mucho más eficiente.

En concreto la tendencia parece indicar que el método 2, donde se ha diseñado una nueva función de K-Means que incorpora la ponderación dentro de la distancia es el doble de lento que el otro.

Es por ello que finalmente se ha optado por aplicar el método 1 donde se adaptan los datos anteriormente al cálculo de K-Means.

-**Selección de ponderaciones:** En este segmento del programa se seleccionan los distintos valores de las ponderaciones.

El funcionamiento está basado en dividir las posibles ponderaciones en distintos puntos equidistantes y aplicar el método de K-Means para cada una de estas ponderaciones, esta iteración de K-Means se haría de acuerdo con lo especificado anteriormente.

En una primera versión, para cubrir todas las casuísticas distintas, se ha hecho uso de funciones de tipo “for”.

En este método se crean diferentes bucles, en cada uno de ellos primero se fija la ponderación que debe ser igual a 1 y a continuación se recorren todas las dife-

rentes posibles ponderaciones que se pueden dar cuando dicha variable tiene una ponderación igual a 1.

A modo de ejemplo:

Sea un vector ponderación que incorporaría las distintas ponderaciones y un valor de precisión que determina en que intervalos se divide el vector ponderación.

Entonces el ciclo correspondiente para las ponderaciones con la primera componente igual a uno seria:

```
  Pond[1]<-1
  for(i in 1:Precision){
    Pond[2]<-i/Precision
    for(j in 1:Precision){
      Pond[3]<-j/Precision
      for(k in 1:Precision){
        Pond[4]<-k/Precision
        for(l in 1:Precision){
          Pond[5]<-l/Precision
          #En este punto se efect an los c lculos correspondientes
   al analisis de K-Means
          cl<- kmeans(M, 5, nstart = 25)
          v<-cl$cluster

          Epond[numpond,]<-v
          numpond=numpond+1
        }
      }
    }
  }
}
```

El código modelo resultante de aplicar este método se pude ver en anexo 1 parte 5.

Este método sin embargo resulto ser altamente ineficiente, tardando tiempos superiores al minuto para precisiones iguales a 6.

Finalmente, se ha decidido hacer un metido alternativo de calculo especificado en anexo 1 parte 1, donde se puede encontrar el programa empleado definitivamente para el análisis de clustring.

Este método alternativo ejecuta el mismo procedimiento pero está diseñado con el objetivo de reducir el número de funciones tipo for() utilizadas. Estas funciones son uno de los inconvenientes de R, puesto que ralentizan altamente el tiempo de ejecución.

En su lugar se ha optado por utilizar funciones tipo apply() que son la alternativa propuesta por el programa para tratar bases de datos donde hay que realizar una misma operación con elementos consecutivos.

Como resultado de la modificación del programa se logra reducir el tiempo computacional del mismo. Permitiendo, a modo de ejemplo, que las ejecuciones resultantes de dividir los intervalos de ponderaciones en 6 unidades, tarden el mismo tiempo (30 segundos) que las ejecuciones resultantes de dividirlo los intervalos

de ponderaciones en 5 unidades con el método anterior.

-**Elección del clúster resultante**: En este segmento del programa se analizan los datos obtenidos para seleccionar cual es el clúster resultante del programa.

Para obtener el clúster resultante del método, se han tabulado los resultados obtenidos tras cada aplicación de K-Means con las distintas ponderaciones y se ha observado cual era el clúster que se repetía un mayor número de veces.

Se ha optado por el mayor número de repeticiones en lugar de calcular el volumen de cada clúster por cuestiones operativas, puesto que es mucho más veloz de implementar.

13. Resultados

Con la ejecución del programa especificado en el anexo 1 parte 1, se ha aplicado el método de la partición más estable a las empresas pertenecientes al IBEX, Dow jones y Euro Stoxx. Se ha utilizado una precisión resultante de dividir cada segmento de ponderaciones en 10 puntos diferentes pudiendo ser cada $w_i \in \{0, \frac{1}{9}, \frac{2}{9}, \frac{3}{9}, \frac{4}{9}, \frac{5}{9}, \frac{6}{9}, \frac{7}{9}, \frac{8}{9}, 1\}$ El tiempo total de ejecución ha ascendido a 1 minuto 2 segundos.

Tras la aplicación del método de han obtenido los siguientes clústeres:

Para el primer método los presentes en el cuadro 9.

	Empresas
Clúster 1	ACX AENA.MC MMM AAPL CAT HD MCD ALV.DE ASML.AS
Clúster 2	CLNX.MC TRE.MC KO JNJ
Clúster 3	AXP CVX IBM JPM MSFT TRV UTX V WMT BAYN.DE SAF.PA OR.PA BMW.DE AIR.PA SIE.MI
Clúster 4	PG
Clúster 5	BA GS UNH MC.PA VOW.DE
Clúster 6	ANA ACS AMS MTS.MC SAB.MC SAN.MC BKIA.MC BKT.MC BBVA.MC CABK.MC DIA.MC ENG.MC ELE.MC FER.MC GAS.MC GRF.MC IAG.MC IBE.MC ITX.MC IDR.MC COL.MC MAP.MC TL5.MC MEL.MC MRL.MC REE.MC REP.MC SGRE.MC TEF.MC VIS.MC CSCO DWDP XOM GE INTC MRK NKE PFE VZ DIS AI.PA ORA.PA ABI.BR PHIA.AS SAN.PA DBK.DE BNP.PA INGA.AS SU.PA DTE.DE BN.PA ENGI.PA G.MI ENEL.MI FRE.DE ENI.MI DPW.DE VIV.PA UNA.AS

Cuadro 9: Clústers método partición más estable

En paralelo se pude comparar con los resultados obtenidos de aplicar el método de K-Means tradicional, utilizando únicamente la distancia euclidea. Los resultados son similares y se pueden observar en el cuadro 10.

A pesar del elevado grado de similitud, se puede observar que existen variaciones entre ambos métodos. En concreto la empresa ACS se clasifica en el clúster 6 según el método de la partición más estable mientras que con K-Means se clasifica en el clúster 3.

Aun así se observa que ambos resultados son prácticamente idénticos. Se conjetura que la causa de este parecido esta en el hecho que los datos muestran todas las variables en distintas escalas.

Este hecho podría suponer un problema para el método de la partición más estable ya que podría suceder que el efecto de tomar todas las ponderaciones se viese

	Empresas
Clúster 1	ACX AENA.MC MMM AAPL CAT HD MCD ALV.DE ASML.AS
Clúster 2	CLNX.MC TRE.MC KO JNJ
Clúster 3	ACS AXP CVX IBM JPM MSFT TRV UTX V WMT BAYN.DE SAF.PA OR.PA BMW.DE AIR.PA SIE.MI
Clúster 4	PG
Clúster 5	BA GS UNH MC.PA VOW.DE
Clúster 6	ANA AMS MTS.MC SAB.MC SAN.MC BKIA.MC BKT.MC BBVA.MC CABK.MC DIA.MC ENG.MC ELE.MC FER.MC GAS.MC GRF.MC IAG.MC IBE.MC ITX.MC IDR.MC COL.MC MAP.MC TL5.MC MEL.MC MRL.MC REE.MC REP.MC SGRE.MC TEF.MC VIS.MC CSCO DWDP XOM GE INTC MRK NKE PFE VZ DIS AI.PA ORA.PA ABI.BR PHIA.AS SAN.PA DBK.DE BNP.PA INGA.AS SU.PA DTE.DE BN.PA ENGI.PA G.MI ENEL.MI FRE.DE ENI.MI DPW.DE VIV.PA UNA.AS

Cuadro 10: Clústers K-Means

limitado por el hecho que las ponderaciones más alejadas de la distancia euclidea tengan menor representación en el método.

Siendo este el caso, para corregir el efecto de estar analizando variables de diferentes tamaños se debe en primer lugar, llevar todas las variables a la misma escala.

Para realizar esta adaptación se opta por calcular la media aritmética de cada variable y dividir las variables por su media. Escalando de esta forma todas las variables alrededor del número 1. El código resultante se encuentra en el anexo 1 parte 2. Tras la aplicación del método de la PMS adaptado se han obtenido los clústeres descritos en el cuadro 11.

En paralelo se pude comparar con los resultados obtenidos de aplicar el método de K-Means tradicional, utilizando únicamente la distancia euclidea. Los resultados son similares y se muestran en el cuadro 12.

Se puede observar que los resultados en esta ocasión muestra una mayor diferencia entre ambos métodos. En concreto, el clúster 1, 2 y 4 se mantienen, mientras los clústeres 3, 5 y parte del 6 se han reorganizado en gran medida.

Observamos pues el efecto de evaluar ponderaciones diferentes a la euclidea, que nos permite obtener clústeres similares aunque no iguales. Estos nuevos clústeres serán los resultantes para un mayor número de ponderaciones diferentes que no únicamente para la distancia euclidea.

	Empresas
Clúster 1	ANA ACS
Clúster 2	MAP.MC
Clúster 3	ACX AENA.MC MMM AAPL CAT HD MCD ALV.DE ASML.AS
Clúster 4	PG
Clúster 5	BA GS UNH MC.PA VOW.DE
Clúster 6	AMS MTS.MC SAB.MC SAN.MC BKIA.MC BKT.MC BBVA.MC CABK.MC CLNX.MC DIA.MC ENG.MC ELE.MC FER.MC GAS.MC GRF.MC IAG.MC IBE.MC ITX.MC IDR.MC COL.MC TL5.MC MEL.MC MRL.MC REE.MC REP.MC SGRE.MC TRE.MC TEF.MC VIS.MC AXP CVX CSCO KO DWDP XOM GE IBM INTC JNJ JPM MRK MSFT NKE PFE TRV UTX VZ V WMT DIS AI.PA ORA.PA ABI.BR PHIA.AS BAYN.DE SAF.PA SAN.PA DBK.DE BNP.PA INGA.AS SU.PA DTE.DE BN.PA OR.PA BMW.DE ENGI.PA AIR.PA G.MI ENEL.MI FRE.DE ENI.MI DPW.DE VIV.PA UNA.AS SIE.MI

Cuadro 11: Clústers método partición más estable variables reescaladas

	Empresas
Clúster 1	ANA ACS
Clúster 2	MAP.MC
Clúster 3	AENA.MC MMM BA GS HD UNH MC.PA VOW.DE
Clúster 4	PG
Clúster 5	CLNX.MC TRE.MC CAT KO JNJ
Clúster 6	ACX AMS MTS.MC SAB.MC SAN.MC BKIA.MC BKT.MC BBVA.MC CABK.MC DIA.MC ENG.MC ELE.MC FER.MC GAS.MC GRF.MC IAG.MC IBE.MC ITX.MC IDR.MC COL.MC TL5.MC MEL.MC MRL.MC REE.MC REP.MC SGRE.MC TEF.MC VIS.MC AXP AAPL CVX CSCO DWDP XOM GE IBM INTC JPM MCD MRK MSFT NKE PFE TRV UTX VZ V WMT DIS AI.PA ORA.PA ABI.BR PHIA.AS ALV.DE BAYN.DE SAF.PA SAN.PA DBK.DE BNP.PA INGA.AS SU.PA DTE.DE ASML.AS BN.PA OR.PA BMW.DE ENGI.PA AIR.PA G.MI ENEL.MI FRE.DE ENI.MI DPW.DE VIV.PA UNA.AS SIE.MI

Cuadro 12: Clústers K-Means variables reescaladas

14. Conclusiones

Tras la observación de los resultados de los apartados anteriores se pueden obtener varias conclusiones:

Distribución de las empresas.

Se considera como distribución definitiva del análisis la obtenida por el método de la partición más estable con datos con el tamaño corregido.

Se considera que esta ejecución es la más libre de biases externos de los datos obtenidos. En primer lugar se ha corregido los biases de tamaño al dividir todos los datos entre su media, dejándolos todos en la misma escala sobre 1.

En segundo lugar se han reducido los biases de ponderación. Al estudiar los diferentes ratios es desconocido la influencia real que tiene cada uno de los ratios sobre la imagen de la empresa. Es decir no podemos conocer que ratio es más y por cuanto más representativo de la empresa estudiada que otro ratio.

En circunstancias corrientes esta clasificación de los ratios se realiza de forma arbitraria por el analista ya sea bien asignando unos pesos o bien no asignándolos y de esta forma haciendo que todas tengan la misma importancia.

Al aplicar el método de la partición más estable se reduce la arbitrariedad de esta decisión, considerando las distintas ponderaciones posibles En el cuadro 12 se puede observar cual es la distribución de las empresas en cada clúster según la lista de origen.

Cúster	IBEX	DOW JONES	Euro Stocs
Clúster 1	2	0	0
Clúster 2	1	0	0
Clúster 3	2	5	2
Clúster 4	0	1	0
Clúster 5	0	3	2
Clúster 6	29	21	25

Cuadro 13: Número empresas por clúster y mercado

Esta distribución, una vez convertidos a porcentajes de empresas de una misma bolsa que hay en cada clúster, podemos ver que resultante es el presente en el cuadro 14

Si se ignoran los clústeres formados por una única empresa, se puede ver que las empresas Americanas del Dow Jones se distancian ligeramente de las empresas Europeas, que incluyen tanto del Euro Stoxx como las del Ibex. I a la vez estos dos conjuntos tienen a comportarse de forma más uniforme.

Aun así esta diferencia es muy leve y a excepción del clúster 3 en el resto de clústeres la distribución es equitativa entre las tres grandes bolsas. Se pude deducir que en un plazo de tiempo de tres años las distintas bolsas siguen ciclos similares y las distintas variables que presentan tienden a ser similares independientemente de su localización.

Cúster	IBEX	DOW JONES	Euro Stoxx
Clúster 1	1	0	0
Clúster 2	1	0	0
Clúster 3	0,22	0,56	0,22
Clúster 4	0	1	0
Clúster 5	0	0,6	0,4
Clúster 6	0,39	0,28	0,33

Cuadro 14: Porcentaje empresas por clúster y mercado

Se supone que esto es un efecto de la globalización que termina causando en última instancia que las características de la empresa no dependan del lugar donde está ubicada su sede social y las políticas de dicho país sino de características globales más generales.

Como diversificar si eres un inversor

En el cuadro 15 se pueden ver las características de cada una empresas y cuál es el valor medio de cada una de las variables de dicho clúster.

Clúster	Rend	Vol	Payout	MB	PE
1	-48,04	0,31	0,6	0,57	0,12
2	-0,0055	0,001	46,07	0,41	0,063
3	0,09	3,24	0,44	1,19	1,04
4	0,004	0,21	0,50	0	58,59
5	0,088	8,50	0,19	0,74	1,93
6	0,025	0,27	0,53	1,02	0,20

Cuadro 15: Media de variables por clúster

Se puede ver que los clústeres 1,2 y 4 se corresponden a empresas con valores excepcionalmente elevados en ciertas características que han formado sus propios clústeres. Se considerarían como casos excepcionales, dichas empresas probablemente no vayan a mantener estos valores puesto que no son sostenibles en el tiempo. Un inversor debería por lo tanto evitar estos valores.

Por otro lado se puede observar que para los clústeres 3,5 y 6 su mayor variable diferencial es la volatilidad. Las empresas se han agrupado según el riesgo que suponen, mientras que la distribución de las demás variables es más o menos homogénea. Así mismo se puede observar también el fenómeno comentado anteriormente, con una mayor volatilidad, es decir a mayor riesgo, se produce un incremento de los rendimientos medios que se obtienen.

Por lo tanto, de cara a diversificar, un inversor deberá evaluar su perfil de riesgo y seleccionar empresas de los clústeres en función de este mismo. Si por ejemplo fuese un inversor conservador y quisiese distribuir su inversión en un 70 % en valores seguros, un 20 % en valores algo más arriesgados y un 10 % buscando beneficio sin tener en cuenta el riesgo, entonces debería destinar el 70 % de la inversión al clúster 6, el 20 % al clúster 3 y el 10 % restante al clúster 5.

Rendimiento del análisis.

Tras la ejecución de ambos análisis se puede observar que el tiempo de ejecución del método de la partición más considerablemente más lento que aplicar K-Means tradicional. En concreto K-Means ejecuta el análisis completo incluyendo la captación y el tratamiento de los datos en 23 segundos. Aplicando el método PMS para 5 variables y analizando 10 valores de ponderación por variable, se tarda 1 minuto 10 segundos en realizar el mismo análisis.

En caso de que se decidiese incrementar la precisión a 15 valores por cada ponderación y variable, el tiempo de ejecución ascendería a 3 minutos y 52 segundos.

Se puede observar por lo tanto un crecimiento exponencial del tiempo a medida que se decide incrementar la precisión, tal y como se había visto en apartados anteriores, ya que debe ejecutarse el análisis de K-Means un gran número de veces.

Esta gran diferencia en tiempo de ejecución, es el principal inconveniente del método y probablemente la razón por la que no se haya intentado investigar en este sentido. No resulta sorprendente que en el momento que surgió el análisis de clústering y cuando no existían herramientas informáticas, nadie quisiese aplicar un método que fácilmente podría escalar el tiempo de análisis miles o millones de veces.

Sin embargo, haciendo uso de la tecnología actual, es posible llevar el análisis de la partición más estable a unos tiempos asequibles para el analista. Quien debido a la naturaleza del análisis podría estar interesado en invertir más tiempo en el análisis con la finalidad de corregir el efecto de asignar ponderaciones de forma arbitraria.

Una de las ventajas del método que hay que volver a destacar es que no es exclusivo para kmeans. El método de PMS se puede utilizar como complemento a la mayoría de métodos de análisis de clúster basados en distancias. Siendo así mucho más versátil en su alcance de aplicación.

Comparativa respecto a K-Means tradicional

De los análisis realizados se puede observar que no hay grandes variaciones entre el método de la partición más estable y el método de K-Means tradicional.

Por un lado se puede observar que ambos métodos dan resultados similares en ambas ejecuciones. Este resultado en si no es sorprendente. Se observa que los resultados fel método PMS no es radicalmente diferente a K-Means, en todo caso al contrario, son similares.

En primer lugar, esta similitud se puede explicar puesto que el método PMS ejecuta el cálculo de K-Means reiteradamente y contrasta diferentes resultados de K-Means para ver cuál es el que se da en más casos. Por lo tanto el clúster resultante de este método es el clúster resultante no solo para una versión de K-Means sino para múltiples de ellas. De ello se puede deducir que será un resultado similar a K-Means tradicional.

Adicionalmente, el método de la partición más estable se define a partir de un espacio de ponderaciones cuya partición central es la ponderación de todas las

distancias por 1. Es decir, la ponderación central es la que corresponde a la distancia euclidea.

Esta forma de definir el espacio de ponderaciones da mayor relevancia a las ponderaciones cercanas a la ponderación central respecto a las más alejadas.

Por ejemplo el espacio que ocupan las ponderaciones de una variable en el intervalo desde que tiene el mismo peso que otra hasta que tiene el doble de importancia, representa el intervalo definido por $[\frac{1}{2}, 1]$ con una longitud de $\frac{1}{2}$.

Mientras el área ocupada por las ponderaciones de la variable desde que representa el doble de peso hasta que representa cuatro veces más, es el intervalo $[\frac{1}{4}, \frac{1}{4}]$ con una longitud de $\frac{1}{4}$. Este conjunto de ponderaciones tendría la mitad de representación que el anterior.

En segundo lugar una parte importante de la similitud entre ambos resultados viene derivada del hecho que ambos análisis se han efectuado sobre el mismo conjunto de datos. Cabe recordar que los métodos de clústering se realizan para clasificar datos similares y relacionarlos entre ellos.

La relación y la similitud entre unos elementos y otros son independientes al método de clústering y se utiliza el análisis para localizar estos puntos en común. Por lo tanto, se intuye que diversos métodos de clústering, a pesar de utilizar diferentes procedimientos, deberán dar clústeres resultantes coherentemente similares.

Un segundo efecto a observar en ambos métodos es que su similitud depende del factor escala. Se puede ver que en el testeo de métodos, que la primera prueba que se ha realizado ha dado como resultado clústeres mucho más similares entre ambos métodos.

Este resultado viene derivado del efecto explicado anteriormente, que en el método de la partición más estable las ponderaciones más alejadas de la central tienen menor representación que aquellas más centrales.

Esto implica que en escenarios en los cuales las variables se encuentren en escalas muy diferentes, la de mayor tamaño continuara dominando sobre las de menor tamaño. Puesto que aunque se tengan en cuenta las ponderaciones que equiparan ambas variables o incluso invierten el tamaño relativo, estas tendrán una importancia menor en el método.

Como resultado, para variables dispares, el método de la partición más estable tiende al mismo análisis sesgado que K-Means de representar los clústeres según únicamente la o las variables de mayor tamaño.

Se puede concluir por lo tanto que el método de la partición más estable, cumple con el objetivo de ofrecer una valoración basada en múltiples ponderaciones, solventando el problema de seleccionar una ponderación para las variables de forma arbitraria.

Sin embargo, el análisis de clúster sigue requiriendo de un tratamiento previo de la muestra que equilibre el tamaño de las diferentes variables.

Referencias

[1] Hartigan, John A *Clustering algorithms*, United States of America, 1975.

[2] Arratia, Argimiro *Computational Finance: An Introductory Course with R* , España, 2014

[3] Damodaran, A. *Investment Valuation: Tools and Techniques for Determining the Value of Any Asset*, New York,United States of America , 2012

[4] Khan,A. y Zuberi, V. *Stock Investing for Everyone: Tools for Investing Like the Pros.* , Hoboken, United States of America, 1999

[5] Marin Ferrer, José Maria *Análisis de Cluster y Arboles de Clasificación [online]* , Universidad Complutense de Madrid, Available at: http://halweb.uc3m.es/esp/Personal/personas/jmmarin/esp/DM/tema6dm.pdf [Accessed 22 Feb 2018]

[6] Universidad de Valencia *Introducción al análisis cluster[online]* , Available at: https://uv.es/ceaces/multivari/cluster/CLUSTER2.htm [Accessed 22 Feb 2018]

[7] R-Project *Contributed Packages[online]* , Available at: https://cran.r-project.org/web/packages/ [Accessed 14 Abril 2018]

15. Anexo 1-Programas

15.1. Método partición mas estable

```
library(quantmod)
library(cluster)
library(combinat)
library(gtools)

setwd('~/TFG')

timestart<-Sys.time()

M<-matrix(sample(1:30,93,replace=T),93,Variables)
Variables<-5
Precision<-9
Empresas<-93
Pond<-numeric(Variables)


EmpresasIBEX<-c("ANA","ACX","ACS","AENA.MC","AMS","MTS.MC","SAB.MC","
    SAN.MC","BKIA.MC","BKT.MC","BBVA.MC","CABK.MC","CLNX.MC","DIA.MC","
    ENG.MC","ELE.MC","FER.MC","GAS.MC","GRF.MC","IAG.MC","IBE.MC","ITX.
    MC","IDR.MC","COL.MC","MAP.MC","TL5.MC","MEL.MC","MRL.MC","REE.MC",
    "REP.MC","SGRE.MC","TRE.MC","TEF.MC","VIS.MC")
EmpresasDOW<-c("MMM","AXP","AAPL","BA","CAT","CVX","CSCO","KO","DWDP",
    "XOM","GE","GS","HD","IBM","INTC","JNJ","JPM","MCD","MRK","MSFT","
    NKE","PFE","PG","TRV","UNH","UTX","VZ","V","WMT","DIS")
EpresasEuro<-c("AI.PA","ORA.PA","ABI.BR","PHIA.AS","ALV.DE","BAYN.DE",
    "SAF.PA","SAN.PA","DBK.DE","BNP.PA", "INGA.AS","SU.PA","DTE.DE","
    ASML.AS","BN.PA","OR.PA","BMW.DE","ENGI.PA","AIR.PA","MC.PA","G.MI"
    ,"ENEL.MI","FRE.DE","ENI.MI","DPW.DE","VOW.DE","VIV.PA","UNA.AS","
    SIE.MI")

ListaEmpresas<-c(EmpresasIBEX,EmpresasDOW,EmpresasEuro)
Listarend<-seq(1,length(ListaEmpresas))
Listavol<-seq(1,length(ListaEmpresas))


for(j in 1:length(ListaEmpresas)){
     getSymbols(ListaEmpresas[j],src='yahoo',from = "2015-01-01",
                to = "2017-12-31")

  saveSymbols(ListaEmpresas[j],file.path = '~/TFG')

  Infofin<-get(load(paste('~/TFG/',ListaEmpresas[j],'.RData',sep='')))
  Cierres<-na.approx(Infofin[,4])

    z<-matrix(0,nrow=floor(length(Cierres)/5)-1,ncol=1)
  for(i in 1:floor(length(Cierres)/5)-1){
    z[i]<-(as.numeric(Cierres[(i+1)*5])-as.numeric(Cierres[i*5]))/as.
    numeric(Cierres[i]*5)

  }

```

```
    Listarend[j]<-mean(z)*100
42  Listavol[j]<-var(Cierres)

44 }

46

48 DivIBEX<-c
     (0.75,0.03,1.01,0.47,0.41,0,0.36,0.53,0.65,0.46,0.63,0.46,0.62,

     1,0.69,0.97,0.65,1.08,0.32,0.19,0.37,0.46,0,0.09,067,0.74,0.23,
50           0.05,0.71,0.6,0.36,4.58,0.04,0.57)

52 PEIBEX<-c
     (18.54,16.93,14.28,19.37,30.65,6.03,10.43,11.6,19.72,15.31,12.13,

     11.65,156.43,14.83,11.58,13.78,60.21,15.21,27.52,6.07,14.51,26.15,
54
     16.21,5.08,11.94,13.9,17.86,5.15,14.14,11.21,0,241,16.65,22.03)

56 MBIBEX<-c
     (1.15,1.67,2.71,4.46,9.99,0.68,0.70,0.93,0.90,1.63,1.04,0.96,9.84,

     8.11,2.21,2.08,2.64,1.21,4.50,2.09,1.09,6.69,3.18,0.99,0.99,3.23,
58           1.75,0.92,3.27,0.75,1.20,3.57,2.35,3.41)

60

62 DivDOW<-c(0.72,0.39,0.24,0.40,2.46,0.90,0.00,4.84,2.40,0.65,0.00,0.28,
           0.49,0.98,0.48,7.06,0.76,0.71,2.17,1.09,0.70,0.36,0.73,0.38,
           0.27,0.49,0.31,0.19,0.62,0.23)
64
66 PEDOW<-c(30.6,34.01,18.54,24.54,123.27,26.07,14.99,130.8,78.88,18.06,

     0,22.05,27.21,23.82,23.87,274.47,17.61,26.52,71.93,82.63,32.41,
           10.19,0,17.35,23.38,22.39,10.51,34.09,25.72,14.83)
68
70 MBDOW<-c(12.73,  4.70,
     6.59,141.11,6.91,1.64,3.59,9.97,1.63,1.89,2.35,1.33,
           150.62,8.07,3.07,5.96,1.59,-41.78,4.45,8.41,8.78,3.05,920.70,
           1.60,4.47,3.44,4.35,8.9,3.74,3.73)
72

74 DivEuro<-c
     (0.50,1.07,1.03,0.93,0.50,0.72,0.09,0.99,0.00,0.47,0.47,0.51,

     0.81,0.23,0.61,0.49,0.27,1.89,0.36,0.41,0.60,0.75,0.19,0.86,
76           0.49,0.09,0.41,0.65,0.48)

78 PEEuro<-c(21.33,23.82,20.16,20.55,11.73,11.92,8.94,9.98,0,9.29,10.05,
           19.32,18.12,34.04,16.45,32.44,6.53,25.38,27.3,29.5,10.63,
80         12.42,20.84,16.41,14.72,7.16,22.61,22.06,14.97)

82 MBEuro<-c(2.77,1.26,2.48,2.46,1.28,2.33,3.43,0.56,0.48,0.78,1.18,2.00,
           2.29,5.20,3.06,4.19,1.06,0.94,4.81,4.32,0.95,1.56,2.64,1.04,
84         3.83,0.77,1.60,9.43,2.19)
```

```
86


88


90 ListaDiv<-c(DivIBEX,DivDOW,DivEuro)
   ListaPE<-c(PEIBEX,PEDOW,PEEuro)
92 ListaMB<-c(MBIBEX,MBDOW,MBEuro)

94 M[,1]<-Listarend
   M[,2]<-Listavol
96 M[,3]<-ListaDiv
   M[,4]<-ListaPE
98 M[,5]<-ListaMB

100 posiblespond<-seq(0, 1, 1/Precision)

102 Comb <- data.frame(permutations((Precision+1), Variables ,
        posiblespond ,repeats.allowed=TRUE))

104 colnames(Comb) <- c('v1', 'v2', 'v3', 'v4','v5')

106 RealComb<-unique(Comb[which(Comb[,1]==1 | Comb[,2]==1 | Comb[,3]==1 |
        Comb[,4]==1 | Comb[,5]==1),])

108 #ResulKmeans tiene empresas en vertical y ponderaciones en horizontal,
        da el cluster para cada empresa en esa podneracion
    ResulKmeans<-apply(RealComb,1,function(x) kmeans(M%*%diag(sqrt(x)),
        centers=M[c(1,15,30,45,60,75),]%*%diag(sqrt(x)))$cluster)
110 t(ResulKmeans)

112 #clusters  ponderaciones en vertical y empresas como vector de texto,
        colapsadas, da el cluster para cada empresa en esa podneracion
    Clusters<-apply(t(ResulKmeans),1,function(x) paste(x,collapse='-'))
114 Tabla<-table(apply(t(ResulKmeans),1,function(x) paste(x,collapse='-'))
        )


116

    maxrep<-which.max(Tabla)
118 which(Clusters==names(maxrep))
    Clusters[which(Clusters==names(maxrep))]
120


122 #ponderaciones Resul son las ponderaciones que dan lugar al cluster
    PonderacionesResul<-RealComb[which(Clusters==names(maxrep)),]
124 ClusterResul<-t(ResulKmeans)[which(Clusters==names(maxrep))[1],]


126

    Cluster1<-which(ClusterResul==1)
128 Cluster2<-which(ClusterResul==2)
    Cluster3<-which(ClusterResul==3)
130 Cluster4<-which(ClusterResul==4)
    Cluster5<-which(ClusterResul==5)
132 Cluster6<-which(ClusterResul==6)
```

```
134
136 EmpresasCluster1<-ListaEmpresas[Cluster1]
    EmpresasCluster2<-ListaEmpresas[Cluster2]
138 EmpresasCluster3<-ListaEmpresas[Cluster3]
    EmpresasCluster4<-ListaEmpresas[Cluster4]
140 EmpresasCluster5<-ListaEmpresas[Cluster5]
    EmpresasCluster6<-ListaEmpresas[Cluster6]
142

144 IBEX<- rep(1, times=34)
    DOW<- rep(2, times=30)
146 Stocks<- rep(3, times=29)

148 Listamercados<-c(IBEX,DOW,Stocks)

150 Tablamercados<-matrix(0,nrow=6,ncol=3)
    Tablamercados1<-table(Listamercados[Cluster1])
152 Tablamercados2<-table( Listamercados[Cluster2])
    Tablamercados3<-table(Listamercados[Cluster3])
154 Tablamercados4<-table(Listamercados[Cluster4])
    Tablamercados5<-table(Listamercados[Cluster5])
156 Tablamercados6<-table(Listamercados[Cluster6])

158
    timeend<-Sys.time()
```

15.2. Método partición mas estable con varaibles escaladas

```
ibrary(quantmod)
library(cluster)
library(combinat)
library(gtools)
setwd('~/TFG')

timestart<-Sys.time()
M<-matrix(sample(1:30,93,replace=T),93,Variables)
Variables<-5
Precision<-9
Empresas<-93
Pond<-numeric(Variables)

EmpresasIBEX<-c("ANA","ACX","ACS","AENA.MC","AMS","MTS.MC","SAB.MC","
    SAN.MC","BKIA.MC","BKT.MC","BBVA.MC","CABK.MC","CLNX.MC","DIA.MC","
    ENG.MC","ELE.MC","FER.MC","GAS.MC","GRF.MC","IAG.MC","IBE.MC","ITX.
    MC","IDR.MC","COL.MC","MAP.MC","TL5.MC","MEL.MC","MRL.MC","REE.MC",
    "REP.MC","SGRE.MC","TRE.MC","TEF.MC","VIS.MC")

EmpresasDOW<-c("MMM","AXP","AAPL","BA","CAT","CVX","CSCO","KO","DWDP",
    "XOM","GE","GS","HD","IBM","INTC","JNJ","JPM","MCD","MRK","MSFT","
    NKE","PFE","PG","TRV","UNH","UTX","VZ","V","WMT","DIS")

EmpresasEuro<-c("AI.PA","ORA.PA","ABI.BR","PHIA.AS","ALV.DE","BAYN.DE"
    ,"SAF.PA","SAN.PA","DBK.DE","BNP.PA", "INGA.AS","SU.PA","DTE.DE","
    ASML.AS","BN.PA","OR.PA","BMW.DE","ENGI.PA","AIR.PA","MC.PA","G.MI"
    ,"ENEL.MI","FRE.DE","ENI.MI","DPW.DE","VOW.DE","VIV.PA","UNA.AS","
    SIE.MI")

ListaEmpresas<-c(EmpresasIBEX,EmpresasDOW,EmpresasEuro)
Listarend<-seq(1,length(ListaEmpresas))
Listavol<-seq(1,length(ListaEmpresas))
for(j in 1:length(ListaEmpresas)){
  getSymbols(ListaEmpresas[j],src='yahoo',from = "2015-01-01",
             to = "2017-12-31")


  saveSymbols(ListaEmpresas[j],file.path = '~/TFG')
  Infofin<-get(load(paste('~/TFG/',ListaEmpresas[j],'.RData',sep='')))
  Cierres<-na.approx(Infofin[,4])
  z<-matrix(0,nrow=floor(length(Cierres)/5)-1,ncol=1)
  for(i in 1:floor(length(Cierres)/5)-1){
    z[i]<-(as.numeric(Cierres[(i+1)*5])-as.numeric(Cierres[i*5]))/as.
    numeric(Cierres[i]*5)
  }
  Listarend[j]<-mean(z)*100
  Listavol[j]<-var(Cierres)
}


DivIBEX<-c
    (0.75,0.03,1.01,0.47,0.41,0,0.36,0.53,0.65,0.46,0.63,0.46,0.62,
```

```
41
      1,0.69,0.97,0.65,1.08,0.32,0.19,0.37,0.46,0,0.09,067,0.74,0.23,
           0.05,0.71,0.6,0.36,4.58,0.04,0.57)
43 PEIBEX<-c
      (18.54,16.93,14.28,19.37,30.65,6.03,10.43,11.6,19.72,15.31,12.13,

      11.65,156.43,14.83,11.58,13.78,60.21,15.21,27.52,6.07,14.51,26.15,
45
      16.21,5.08,11.94,13.9,17.86,5.15,14.14,11.21,0,241,16.65,22.03)
   MBIBEX<-c
      (1.15,1.67,2.71,4.46,9.99,0.68,0.70,0.93,0.90,1.63,1.04,0.96,9.84,
47
      8.11,2.21,2.08,2.64,1.21,4.50,2.09,1.09,6.69,3.18,0.99,0.99,3.23,
           1.75,0.92,3.27,0.75,1.20,3.57,2.35,3.41)
49


51 DivDOW<-c(0.72,0.39,0.24,0.40,2.46,0.90,0.00,4.84,2.40,0.65,0.00,0.28,
           0.49,0.98,0.48,7.06,0.76,0.71,2.17,1.09,0.70,0.36,0.73,0.38,
53         0.27,0.49,0.31,0.19,0.62,0.23)
   PEDOW<-c(30.6,34.01,18.54,24.54,123.27,26.07,14.99,130.8,78.88,18.06,
55
      0,22.05,27.21,23.82,23.87,274.47,17.61,26.52,71.93,82.63,32.41,
           10.19,0,17.35,23.38,22.39,10.51,34.09,25.72,14.83)
57 MBDOW<-c(12.73,  4.70,
      6.59,141.11,6.91,1.64,3.59,9.97,1.63,1.89,2.35,1.33,
           150.62,8.07,3.07,5.96,1.59,-41.78,4.45,8.41,8.78,3.05,920.70,
59         1.60,4.47,3.44,4.35,8.9,3.74,3.73)


61
   DivEuro<-c
      (0.50,1.07,1.03,0.93,0.50,0.72,0.09,0.99,0.00,0.47,0.47,0.51,
63
      0.81,0.23,0.61,0.49,0.27,1.89,0.36,0.41,0.60,0.75,0.19,0.86,
           0.49,0.09,0.41,0.65,0.48)
65 PEEuro<-c(21.33,23.82,20.16,20.55,11.73,11.92,8.94,9.98,0,9.29,10.05,
           19.32,18.12,34.04,16.45,32.44,6.53,25.38,27.3,29.5,10.63,
67         12.42,20.84,16.41,14.72,7.16,22.61,22.06,14.97)
   MBEuro<-c(2.77,1.26,2.48,2.46,1.28,2.33,3.43,0.56,0.48,0.78,1.18,2.00,
69         2.29,5.20,3.06,4.19,1.06,0.94,4.81,4.32,0.95,1.56,2.64,1.04,
           3.83,0.77,1.60,9.43,2.19)
71
   ListaDiv<-c(DivIBEX,DivDOW,DivEuro)
73 ListaPE<-c(PEIBEX,PEDOW,PEEuro)
   ListaMB<-c(MBIBEX,MBDOW,MBEuro)
75
   M[,1]<-Listarend
77 M[,2]<-Listavol
   M[,3]<-ListaDiv
79 M[,4]<-ListaPE
   M[,5]<-ListaMB
81
   Mediarend<-mean(M[,1])
83 Mediavol<-mean(M[,2])
   Mediadiv<-mean(M[,3])
85 MediaPE<-mean(M[,4])
```

```
MediaMB<-mean(M[,5])

M<-M%*%diag(c(1/Mediarend,1/Mediavol, 1/Mediadiv, 1/MediaPE, 1/MediaMB))
posiblespond<-seq(0, 1, 1/Precision)
Comb <- data.frame(permutations((Precision+1), Variables , posiblespond ,repeats.allowed=TRUE))
colnames(Comb) <- c('v1', 'v2', 'v3', 'v4','v5')
RealComb<-unique(Comb[which(Comb[,1]==1 | Comb[,2]==1 | Comb[,3]==1 | Comb[,4]==1 | Comb[,5]==1),])

#ResulKmeans tiene empresas en vertical y ponderaciones en horizontal, da el cluster para cada empresa en esa podneracion
ResulKmeans<-apply(RealComb,1,function(x) kmeans(M%*%diag(sqrt(x)), centers=M[c(1,15,30,45,60,75),]%*%diag(sqrt(x)))$cluster)
t(ResulKmeans)

#clusters  ponderaciones en vertical y empresas como vector de texto, colapsadas, da el cluster para cada empresa en esa podneracion
Clusters<-apply(t(ResulKmeans),1,function(x) paste(x,collapse='-'))
Tabla<-table(apply(t(ResulKmeans),1,function(x) paste(x,collapse='-')))

maxrep<-which.max(Tabla)
which(Clusters==names(maxrep))
Clusters[which(Clusters==names(maxrep))]

#ponderaciones Resul son las ponderaciones que dan lugar al cluster
PonderacionesResul<-RealComb[which(Clusters==names(maxrep)),]
ClusterResul<-t(ResulKmeans)[which(Clusters==names(maxrep))[1],]
Cluster1<-which(ClusterResul==1)
Cluster2<-which(ClusterResul==2)

Cluster3<-which(ClusterResul==3)
Cluster4<-which(ClusterResul==4)
Cluster5<-which(ClusterResul==5)
Cluster6<-which(ClusterResul==6)

EmpresasCluster1<-ListaEmpresas[Cluster1]
EmpresasCluster2<-ListaEmpresas[Cluster2]
EmpresasCluster3<-ListaEmpresas[Cluster3]
EmpresasCluster4<-ListaEmpresas[Cluster4]
EmpresasCluster5<-ListaEmpresas[Cluster5]
EmpresasCluster6<-ListaEmpresas[Cluster6]

IBEX<- rep(1, times=34)
DOW<- rep(2, times=30)
Stocks<- rep(3, times=29)
Listamercados<-c(IBEX,DOW,Stocks)
Tablamercados<-matrix(0,nrow=6,ncol=3)
Tablamercados1<-table(Listamercados[Cluster1])
Tablamercados2<-table( Listamercados[Cluster2])
Tablamercados3<-table(Listamercados[Cluster3])
Tablamercados4<-table(Listamercados[Cluster4])
Tablamercados5<-table(Listamercados[Cluster5])
```

```
Tablamercados6<-table(Listamercados[Cluster6])
135
137 RendCluser<-c(mean(Listarend[Cluster1]),mean(Listarend[Cluster2]),mean
    (Listarend[Cluster3]),mean(Listarend[Cluster4]),mean(Listarend[
    Cluster5]),mean(Listarend[Cluster6]))
    VolCluser<-c(mean(ListaVol[Cluster1]),mean(ListaVol[Cluster2]),mean(
    ListaVol[Cluster3]),mean(ListaVol[Cluster4]),mean(ListaVol[Cluster5
    ]),mean(ListaVol[Cluster6]))
139 DivCluser<-c(mean(ListaDiv[Cluster1]),mean(ListaDiv[Cluster2]),mean(
    ListaDiv[Cluster3]),mean(ListaDiv[Cluster4]),mean(ListaDiv[Cluster5
    ]),mean(ListaDiv[Cluster6]))
    MBCluser<-c(mean(ListaMB[Cluster1]),mean(ListaMB[Cluster2]),mean(
    ListaMB[Cluster3]),mean(ListaMB[Cluster4]),mean(ListaMB[Cluster5]),
    mean(ListaMB[Cluster6]))
141 PECluser<-c(mean(ListaPE[Cluster1]),mean(ListaPE[Cluster2]),mean(
    ListaPE[Cluster3]),mean(ListaPE[Cluster4]),mean(ListaPE[Cluster5]),
    mean(ListaPE[Cluster6]))

143 timeend<-Sys.time()
```

15.3. Silhouette

```
library(quantmod)
library(cluster)
setwd('~/TFG')

M<-matrix(sample(1:30,93,replace=T),93,Variables)
EmpresasIBEX<-c("ANA","ACX","ACS","AENA.MC","AMS","MTS.MC","SAB.MC","
    SAN.MC","BKIA.MC","BKT.MC","BBVA.MC","CABK.MC","CLNX.MC","DIA.MC","
    ENG.MC","ELE.MC","FER.MC","GAS.MC","GRF.MC","IAG.MC","IBE.MC","ITX.
    MC","IDR.MC","COL.MC","MAP.MC","TL5.MC","MEL.MC","MRL.MC","REE.MC",
    "REP.MC","SGRE.MC","TRE.MC","TEF.MC","VIS.MC")
EmpresasDOW<-c("MMM","AXP","AAPL","BA","CAT","CVX","CSCO","KO","DWDP",
    "XOM","GE","GS","HD","IBM","INTC","JNJ","JPM","MCD","MRK","MSFT","
    NKE","PFE","PG","TRV","UNH","UTX","VZ","V","WMT","DIS")
EmpresasEuro<-c("AI.PA","ORA.PA","ABI.BR","PHIA.AS","ALV.DE","BAYN.DE"
    ,"SAF.PA","SAN.PA","DBK.DE","BNP.PA", "INGA.AS","SU.PA","DTE.DE","
    ASML.AS","BN.PA","OR.PA","BMW.DE","ENGI.PA","AIR.PA","MC.PA","G.MI"
    ,"ENEL.MI","FRE.DE","ENI.MI","DPW.DE","VOW.DE","VIV.PA","UNA.AS","
    SIE.MI")

ListaEmpresas<-c(EmpresasIBEX,EmpresasDOW,EmpresasEuro)

Listarend<-seq(1,length(ListaEmpresas))
Listavol<-seq(1,length(ListaEmpresas))
for(j in 1:length(ListaEmpresas)){
  getSymbols(ListaEmpresas[j],src='yahoo',from = "2015-01-01",
             to = "2017-12-31")
  saveSymbols(ListaEmpresas[j],file.path = '~/TFG')
  Infofin<-get(load(paste('~/TFG/',ListaEmpresas[j],'.RData',sep='')))

  Cierres<-na.approx(Infofin[,4])

  z<-matrix(0,nrow=floor(length(Cierres)/5)-1,ncol=1)
  for(i in 1:floor(length(Cierres)/5)-1){
    z[i]<-(as.numeric(Cierres[(i+1)*5])-as.numeric(Cierres[i*5]))/as.
    numeric(Cierres[i]*5)
  }
  Listarend[j]<-mean(z)*100
  Listavol[j]<-var(Cierres)
}

DivIBEX<-c
    (0.75,0.03,1.01,0.47,0.41,0,0.36,0.53,0.65,0.46,0.63,0.46,0.62,

    1,0.69,0.97,0.65,1.08,0.32,0.19,0.37,0.46,0,0.09,067,0.74,0.23,
            0.05,0.71,0.6,0.36,4.58,0.04,0.57)
PEIBEX<-c
    (18.54,16.93,14.28,19.37,30.65,6.03,10.43,11.6,19.72,15.31,12.13,

    11.65,156.43,14.83,11.58,13.78,60.21,15.21,27.52,6.07,14.51,26.15,

    16.21,5.08,11.94,13.9,17.86,5.15,14.14,11.21,0,241,16.65,22.03)
MBIBEX<-c
    (1.15,1.67,2.71,4.46,9.99,0.68,0.70,0.93,0.90,1.63,1.04,0.96,9.84,
```

```
37
      8.11,2.21,2.08,2.64,1.21,4.50,2.09,1.09,6.69,3.18,0.99,0.99,3.23,
          1.75,0.92,3.27,0.75,1.20,3.57,2.35,3.41)
39
DivDOW<-c(0.72,0.39,0.24,0.40,2.46,0.90,0.00,4.84,2.40,0.65,0.00,0.28,
41          0.49,0.98,0.48,7.06,0.76,0.71,2.17,1.09,0.70,0.36,0.73,0.38,
          0.27,0.49,0.31,0.19,0.62,0.23)
43 PEDOW<-c(30.6,34.01,18.54,24.54,123.27,26.07,14.99,130.8,78.88,18.06,

      0,22.05,27.21,23.82,23.87,274.47,17.61,26.52,71.93,82.63,32.41,
45          10.19,0,17.35,23.38,22.39,10.51,34.09,25.72,14.83)
MBDOW<-c(12.73,  4.70,
      6.59,141.11,6.91,1.64,3.59,9.97,1.63,1.89,2.35,1.33,
47          150.62,8.07,3.07,5.96,1.59,-41.78,4.45,8.41,8.78,3.05,920.70,
          1.60,4.47,3.44,4.35,8.9,3.74,3.73)
49


51 DivEuro<-c
      (0.50,1.07,1.03,0.93,0.50,0.72,0.09,0.99,0.00,0.47,0.47,0.51,

      0.81,0.23,0.61,0.49,0.27,1.89,0.36,0.41,0.60,0.75,0.19,0.86,
53          0.49,0.09,0.41,0.65,0.48)
PEEuro<-c(21.33,23.82,20.16,20.55,11.73,11.92,8.94,9.98,0,9.29,10.05,
55          19.32,18.12,34.04,16.45,32.44,6.53,25.38,27.3,29.5,10.63,
          12.42,20.84,16.41,14.72,7.16,22.61,22.06,14.97)
57 MBEuro<-c(2.77,1.26,2.48,2.46,1.28,2.33,3.43,0.56,0.48,0.78,1.18,2.00,
          2.29,5.20,3.06,4.19,1.06,0.94,4.81,4.32,0.95,1.56,2.64,1.04,
59          3.83,0.77,1.60,9.43,2.19)

61 ListaDiv<-c(DivIBEX,DivDOW,DivEuro)
ListaPE<-c(PEIBEX,PEDOW,PEEuro)
63 ListaMB<-c(MBIBEX,MBDOW,MBEuro)

65 M[,1]<-Listarend
M[,2]<-Listavol
67 M[,3]<-ListaDiv
M[,4]<-ListaPE
69 M[,5]<-ListaMB

71 dis <- dist(M)^2
res <- kmeans(M,20)
73 sil = silhouette (res$cluster, dis)
windows()
75 plot(sil)
```

15.4. K-Means método 2

```
timestart<-Sys.time()
Variables<-5
Precision<-10
Empresas<-4800
Pond<-c(1,1,1,1,1)
Iter<-Variables*Precision^4

M<-matrix(sample(1:1000,24000,replace=T),Empresas,Variables)
numpond<-1
Rep<-numeric(Iter)

K<-5
count<-1
centros<-matrix(0,nrow=K,ncol=Variables)
nelemencluster<-matrix(0,nrow=K,1)
Minc<-0
Mind<-0


# Se crean los primeros clusters de forma artificial, eligiendo grupos
    de empresas/k elementos
Kmeans<-numeric(Empresas)
nelemen<-Empresas%/%K
for(i in 1:(K-1)){
  for(l in 1:nelemen){
    Kmeans[count]<-i
    count=count+1
    centros[i,]=centros[i,]+(M[count,]/nelemen)
  }
}

rep<-0
vueltas<-0
#se asigna cada elemento al cluster mas cercano
while(rep==0&vueltas<10){
  Kmeans2<-Kmeans
  for(i in 1:Empresas){
    Mind<-sum(Pond*(M[i,]-centros[1,])^2)
    Minc<-1
    for(j in 2:K){
      if(sum(Pond*(M[i,]-centros[j,])^2)<Mind){
        Mind<-sum(Pond*(M[i,]-centros[j,])^2)
        Minc<-j
      }
    }
    Kmeans[i]<-Minc
  }
  #Se recalculan los centros de cada cluster
  centros<-matrix(0,nrow=K,ncol=Variables)
  nelemencluster<-matrix(0,nrow=K,1)
  for(i in 1:Empresas){
    clus<-Kmeans[i]
    centros[clus,]<-centros[clus,]+M[i,]
```

```
53      nelemencluster[clus,1]<-nelemencluster[clus,1]+1
    }
55  for(i in 1:K){
      centros[i,]<-centros[i,]/max(nelemencluster[i,1],1)
57  }
  #evalua si se repite el mismo vector
59  rep<-1
  for(i in 1:Empresas){
61    if(Kmeans2[i]!=Kmeans[i]){ rep<-0}
  }
63  vueltas<-vueltas+1
}
65 timeend<-Sys.time()
```

15.5. Ponderación mas estable con bucles

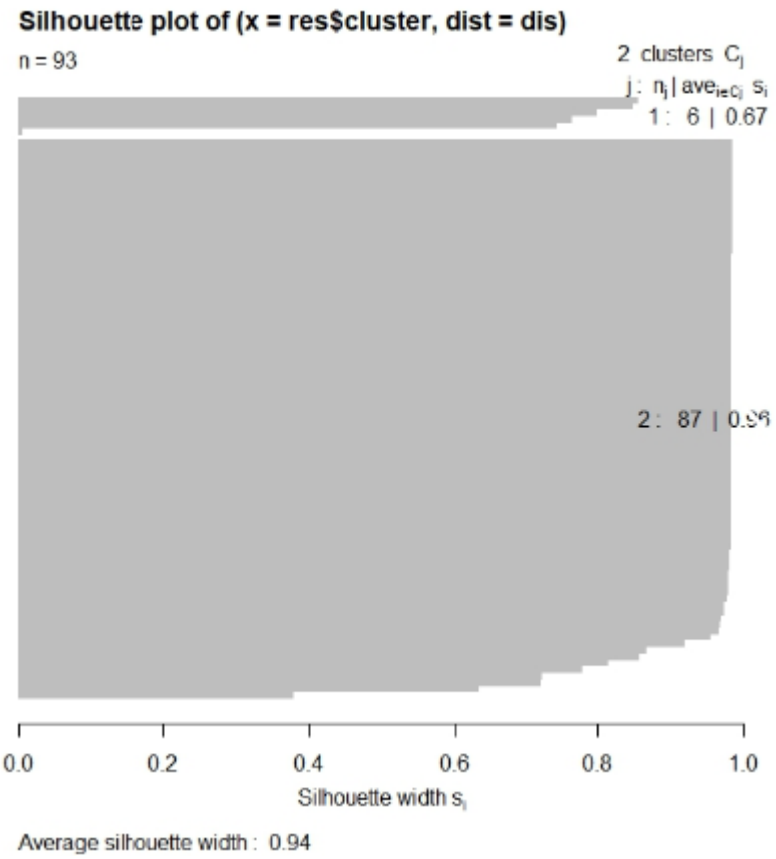
```
timestart<-Sys.time()
Variables<-5
Precision<-5
Empresas<-100
Pond<-numeric(Variables)
Iter<-Variables*Precision^4
M<-matrix(sample(1:30,500,replace=T),Empresas,Variables)
numpond<-1
Epond<-matrix(0,Iter,Empresas)
Rep<-numeric(Iter)


for(k in 1:Variables){
  Pond[1]<-1
  for(i in 1:Precision){
    Pond[2]<-i/Precision
    for(j in 1:Precision){
      Pond[3]<-j/Precision
      for(k in 1:Precision){
        Pond[4]<-k/Precision
        for(l in 1:Precision){
          Pond[5]<-l/Precision
          cl<- kmeans(M, 5, nstart = 25)
          v<-cl$cluster

          Epond[numpond,]<-v
          numpond=numpond+1
        }
      }
    }
  }
}

for(k in 1:Iter){
  for(i in k:Iter){
    if(Epond[k,1]==Epond[i,1]){
      Rep[k]=Rep[k]+1
    }
  }
}
maxrep<-which.max(Rep)
timeend<-Sys.time()
```

16. Anexo 2-Material de apoyo

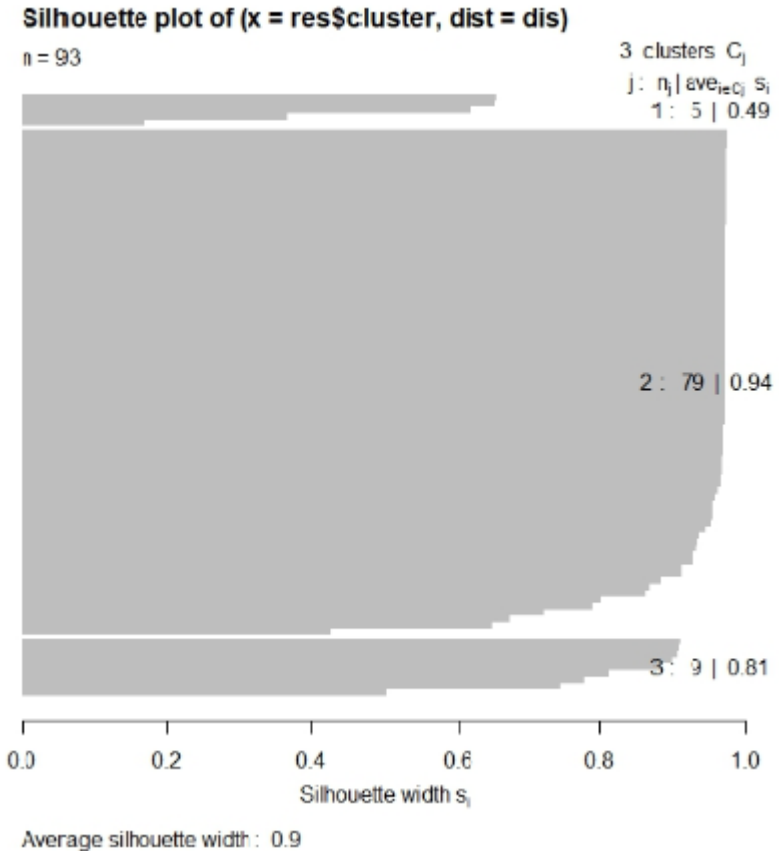
16.1. Resultados silhouette

A continuación se muestran los graficos obtenidos por la función silhouette para los distintos valores posibles de clústeres para los datos originalmente captados.

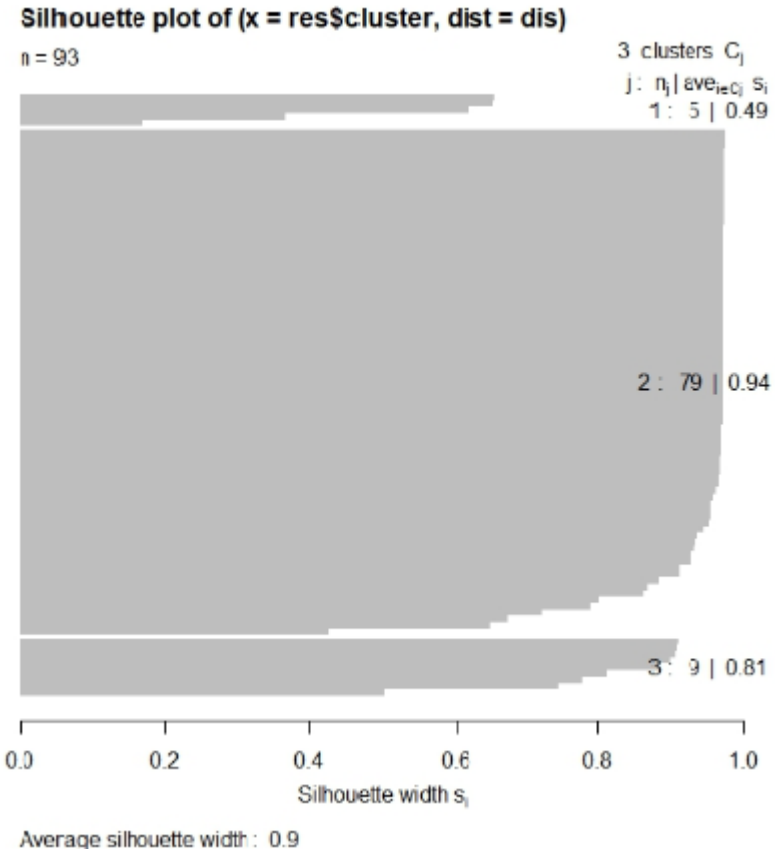
K-Means con K=2



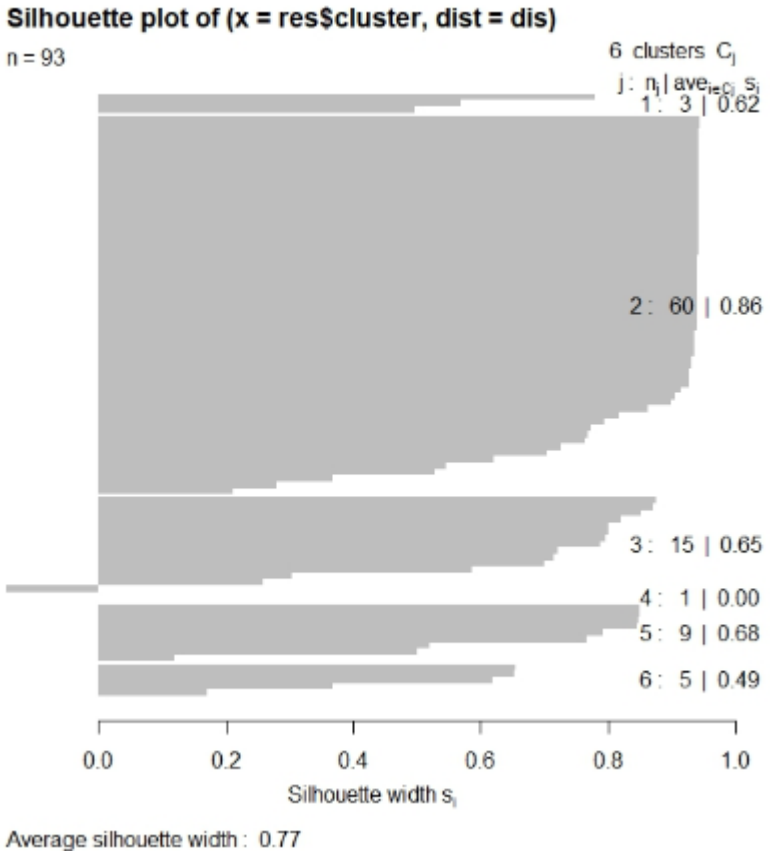
K-Means con K=3



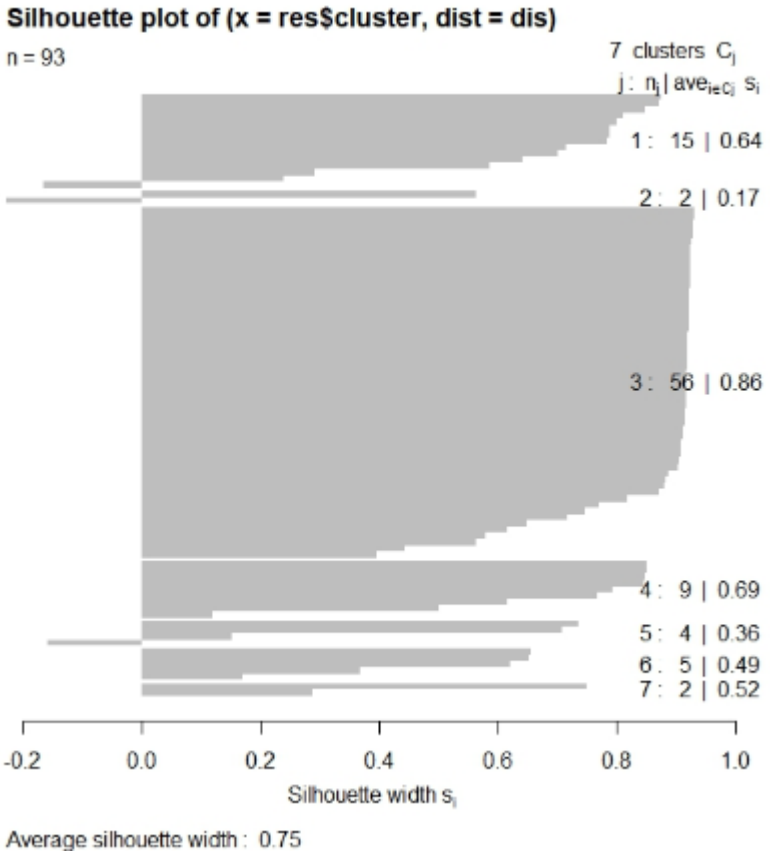
K-Means con K=5



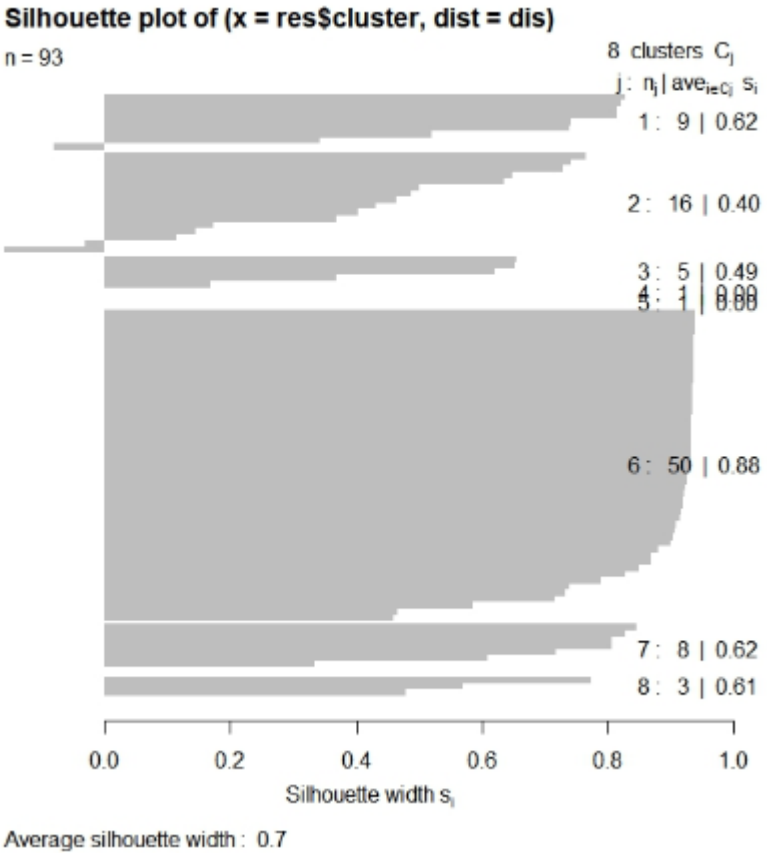
K-Means con K=6



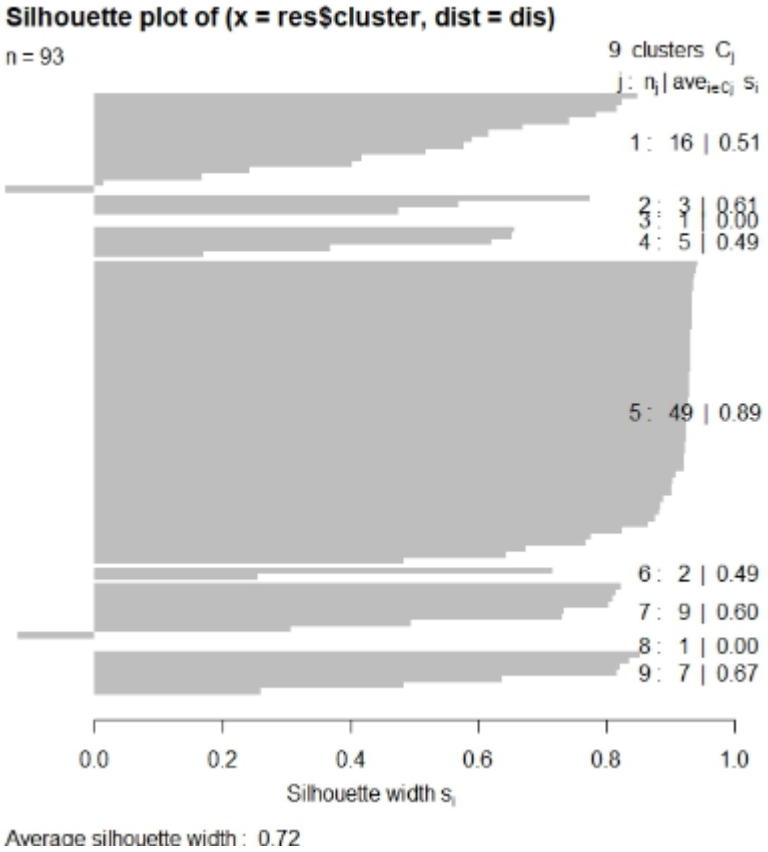
K-Means con K=7



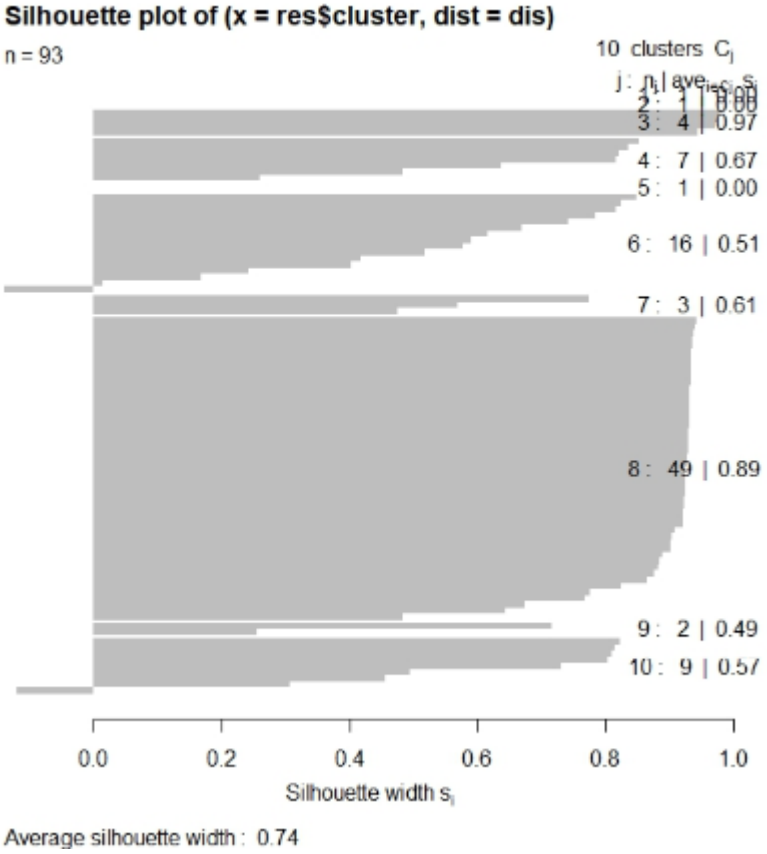
K-Means con K=8



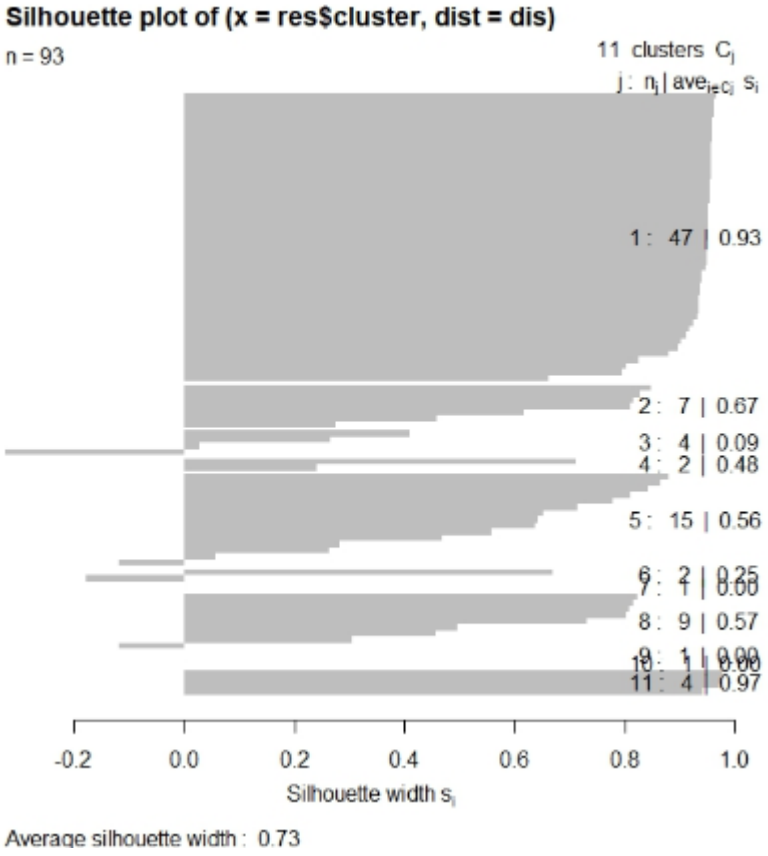
K-Means con K=9



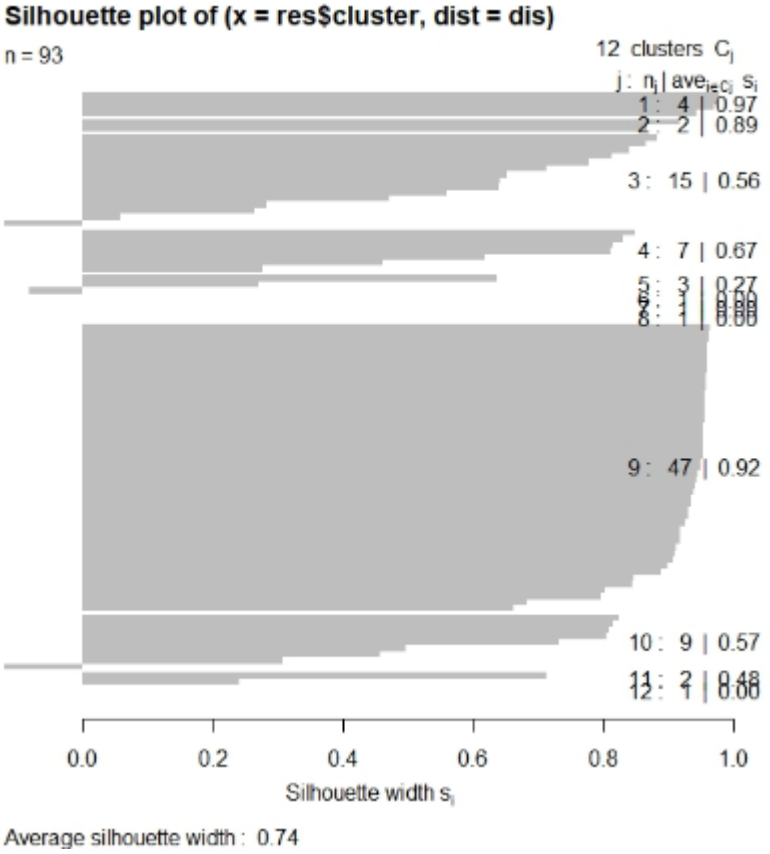
K-Means con K=10



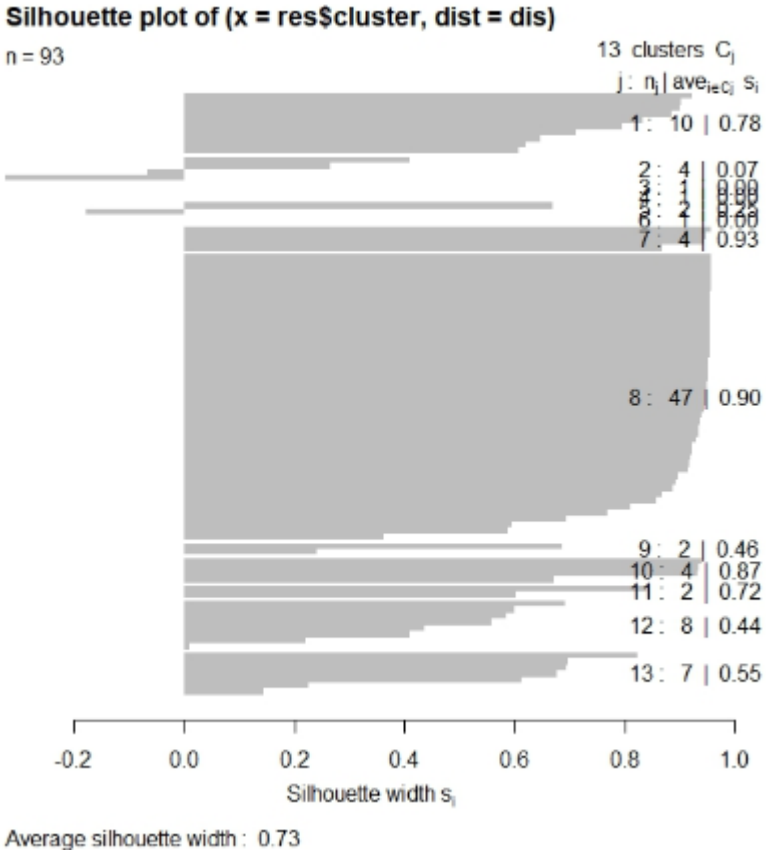
K-Means con K=11



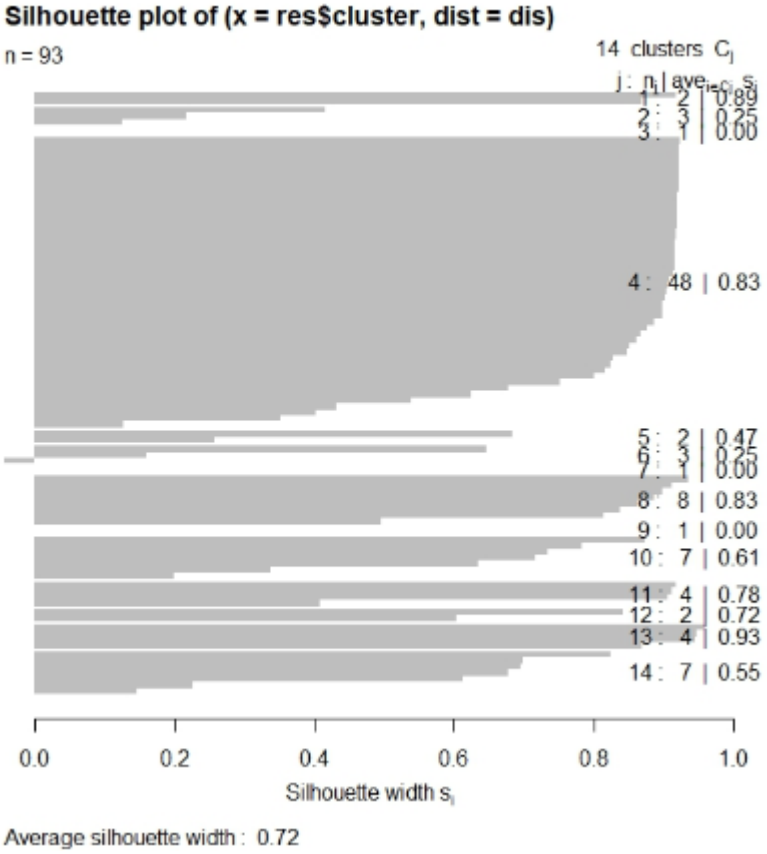
K-Means con K=12



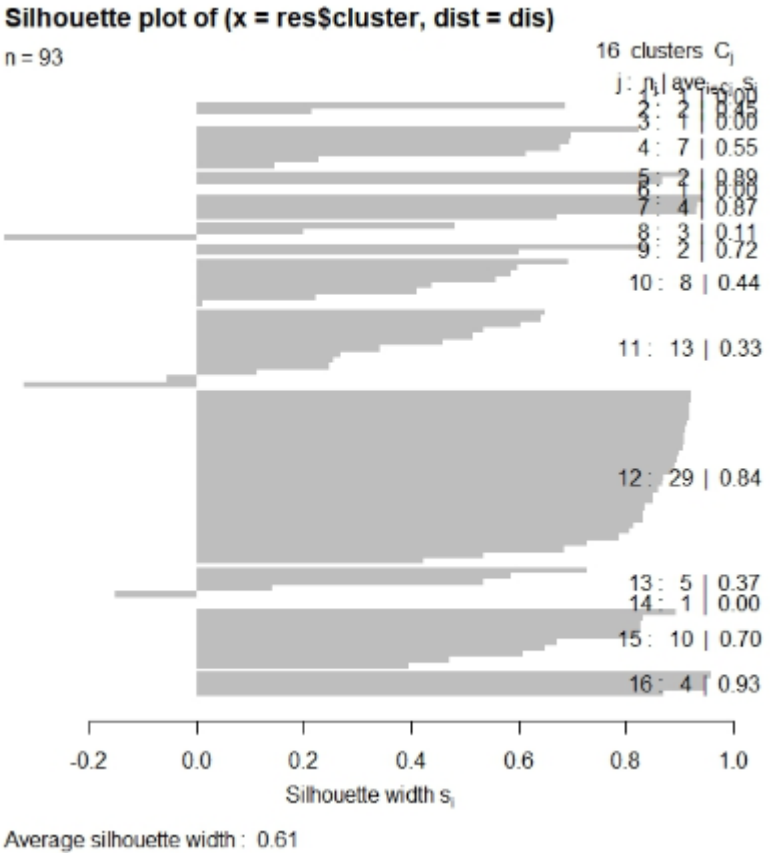
K-Means con K=13



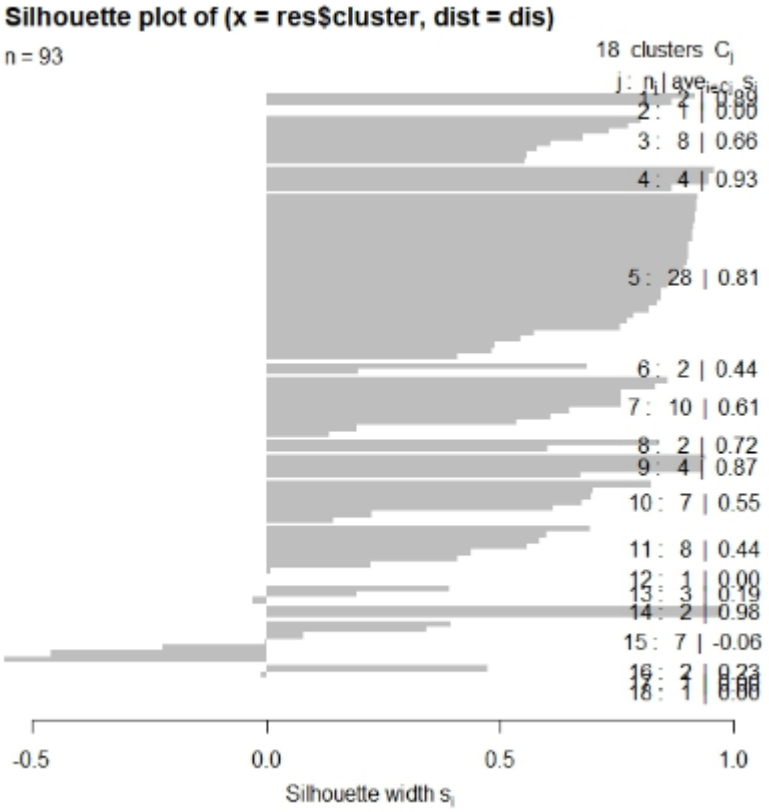
K-Means con K=14



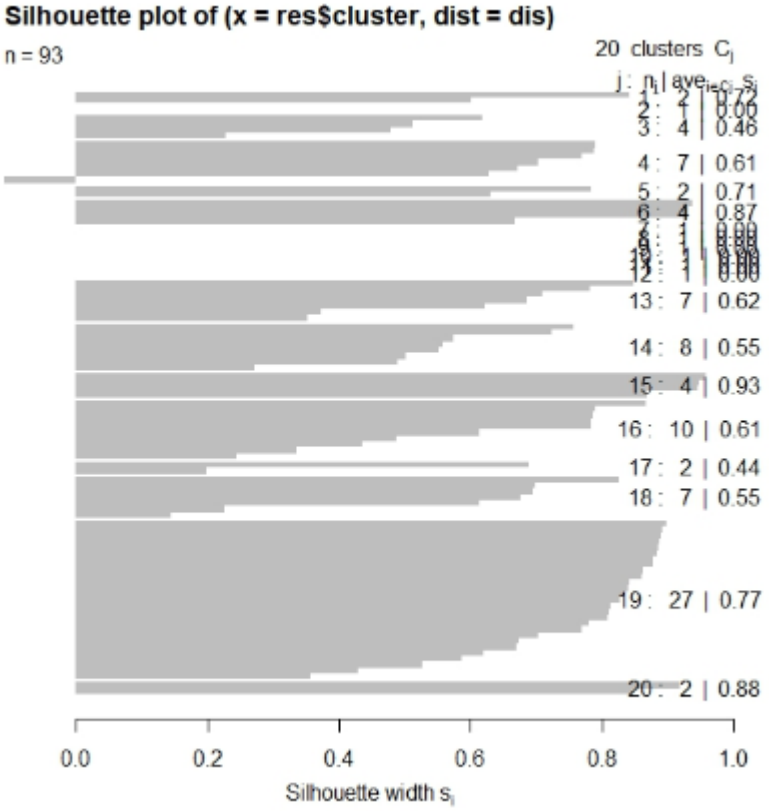
K-Means con K=16



K-Means con K=18



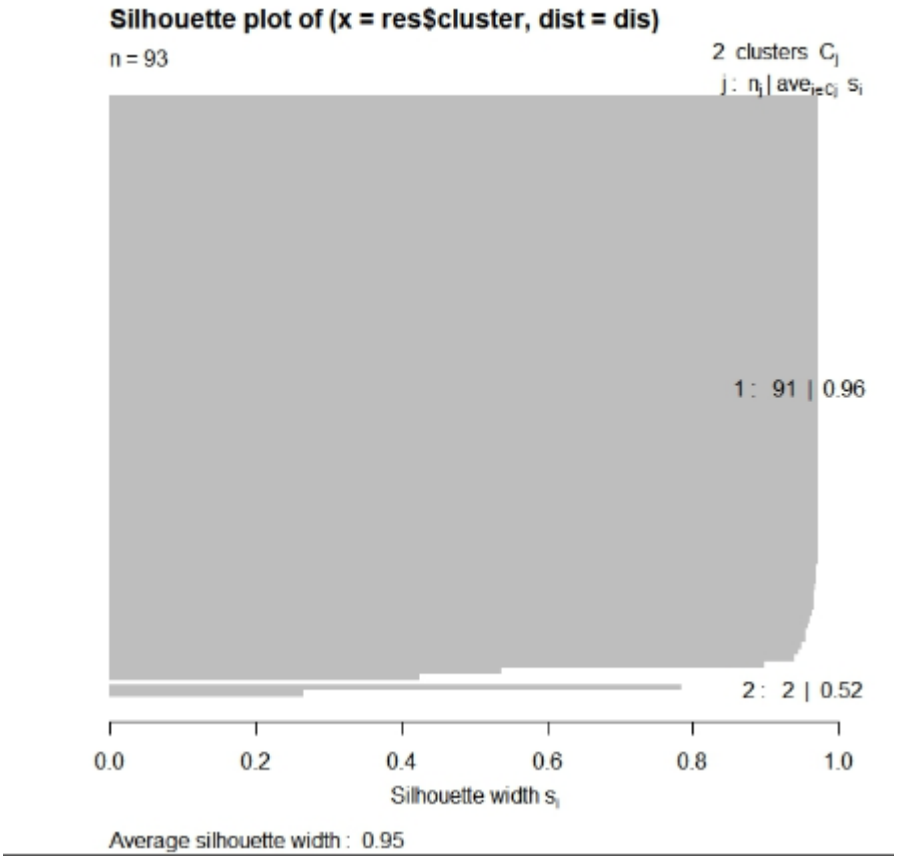
K-Means con K=20



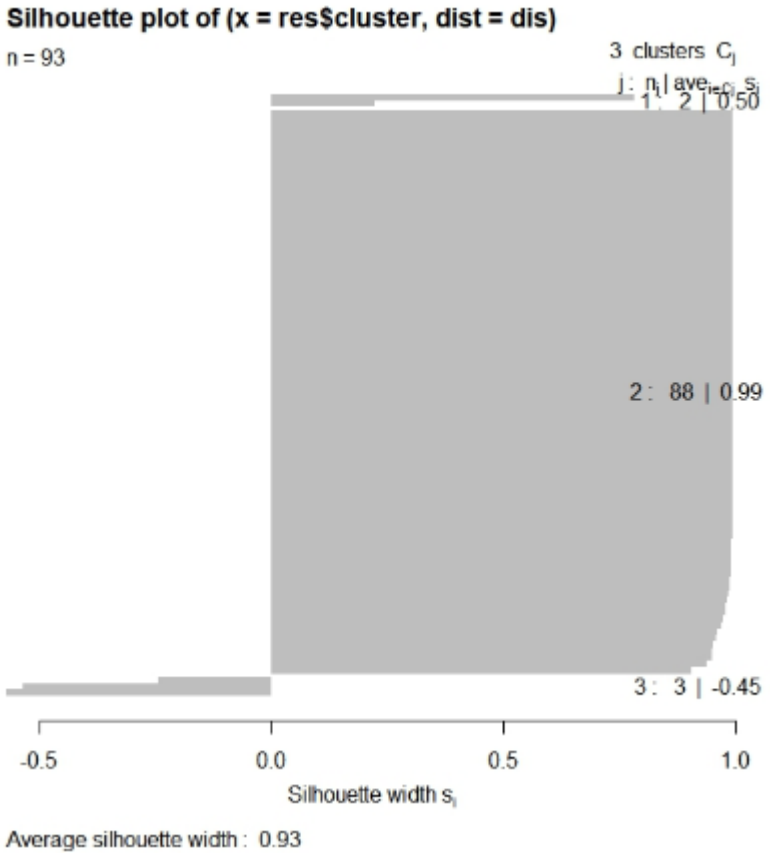
16.2. Resultados silhouette para variables reescaladas

A continuación, en los cuadros del 16 al 21 se muestran los graficos obtenidos por la función silhouette para los distintos valores posibles de clústeres para los datos dispuestos en la misma escala.

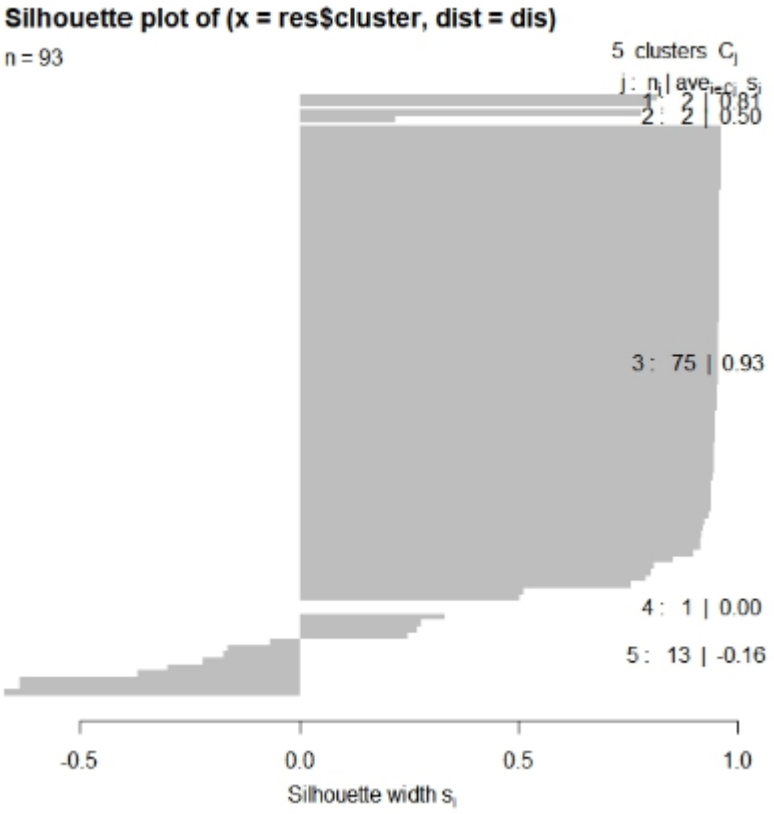
K-Means con K=2



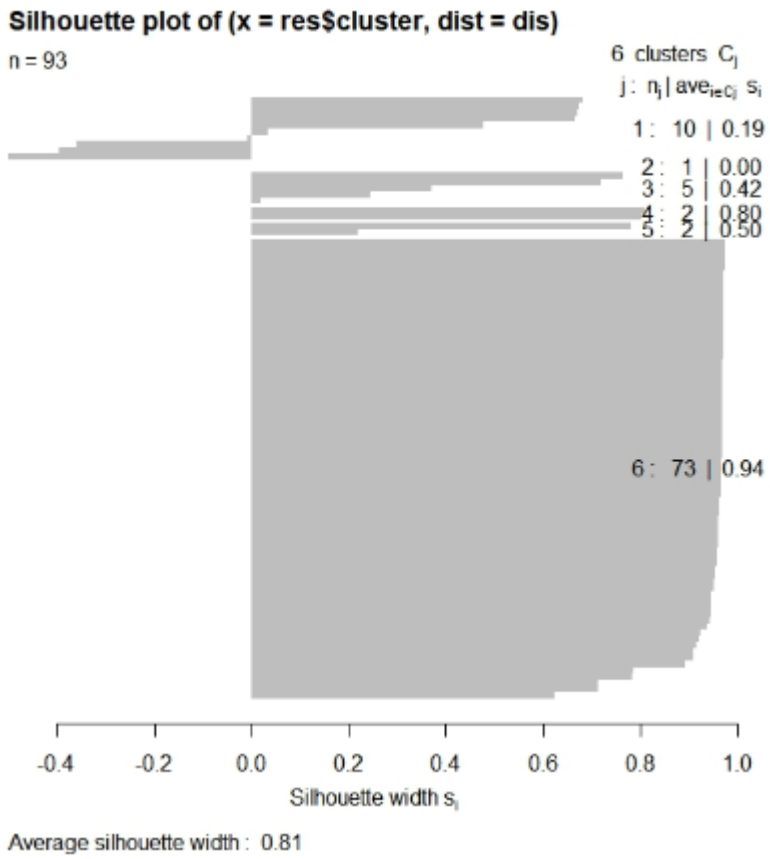
K-Means con K=3



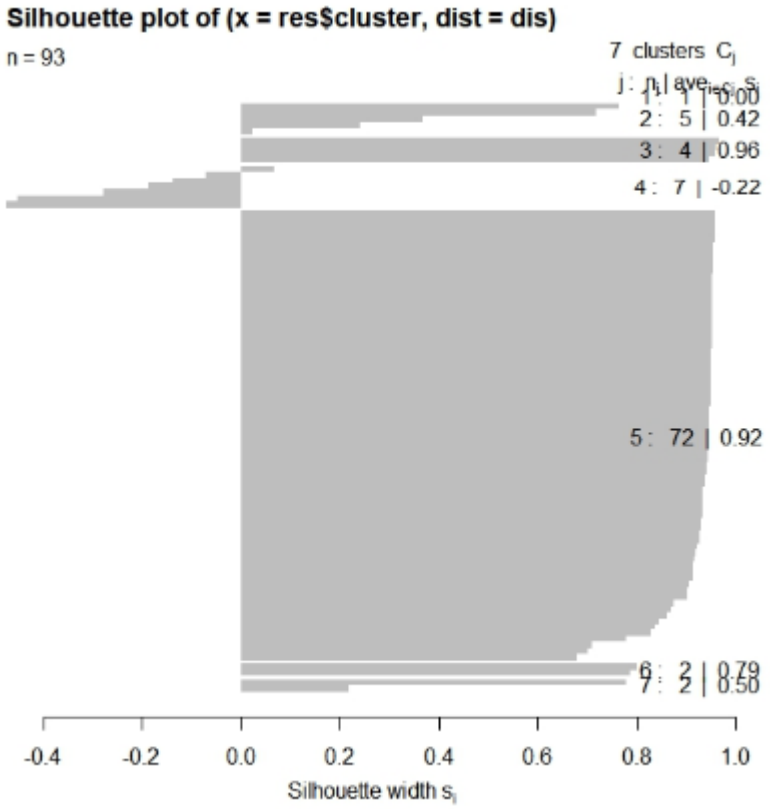
K-Means con K=5



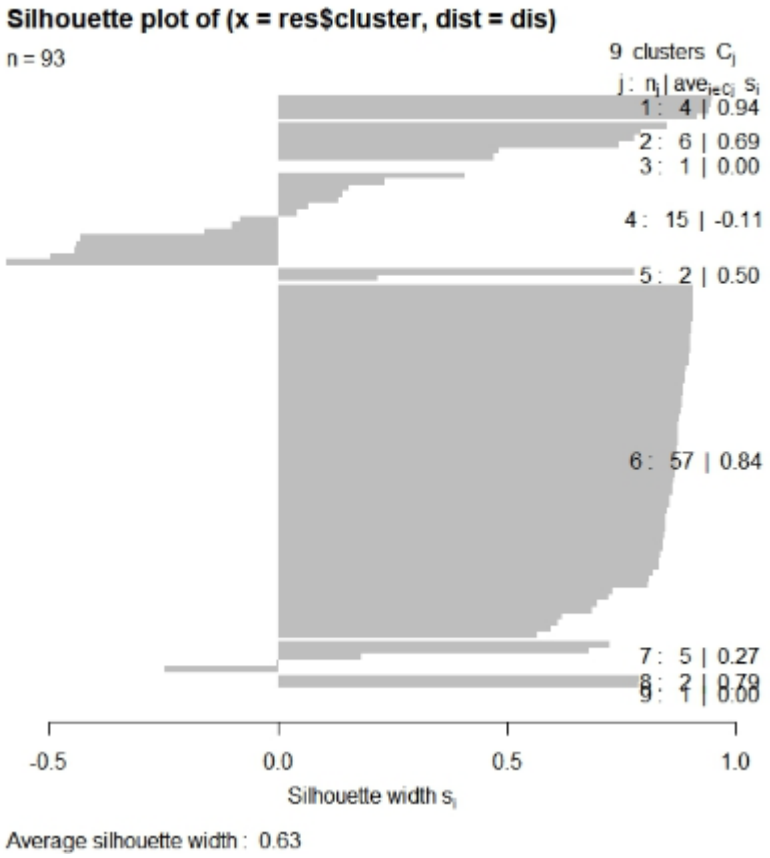
K-Means con K=6



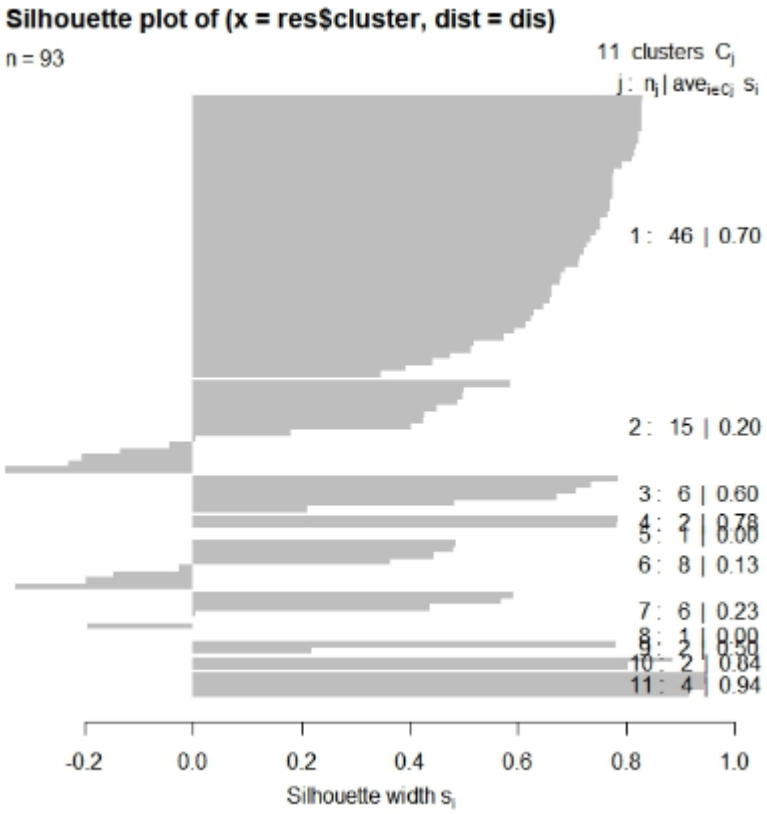
K-Means con K=7



K-Means con K=9

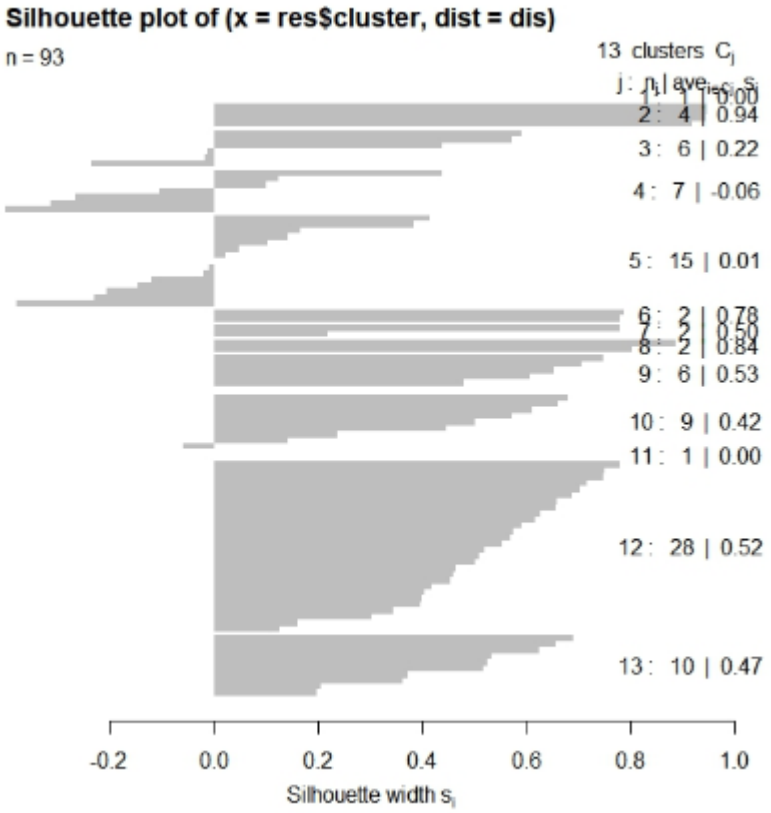


K-Means con K=11



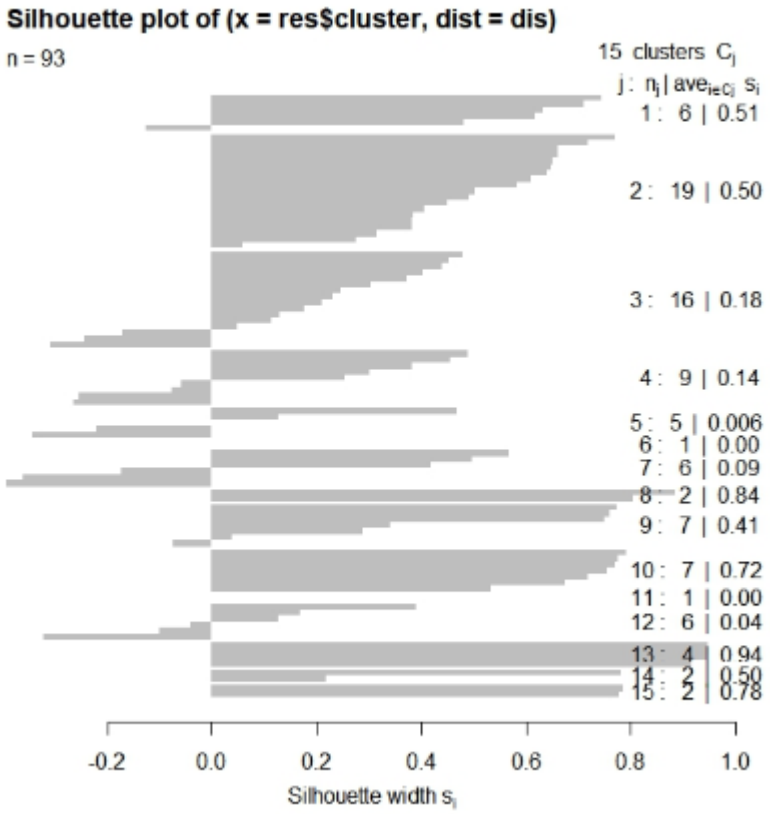
Average silhouette width : 0.53

K-Means con K=13

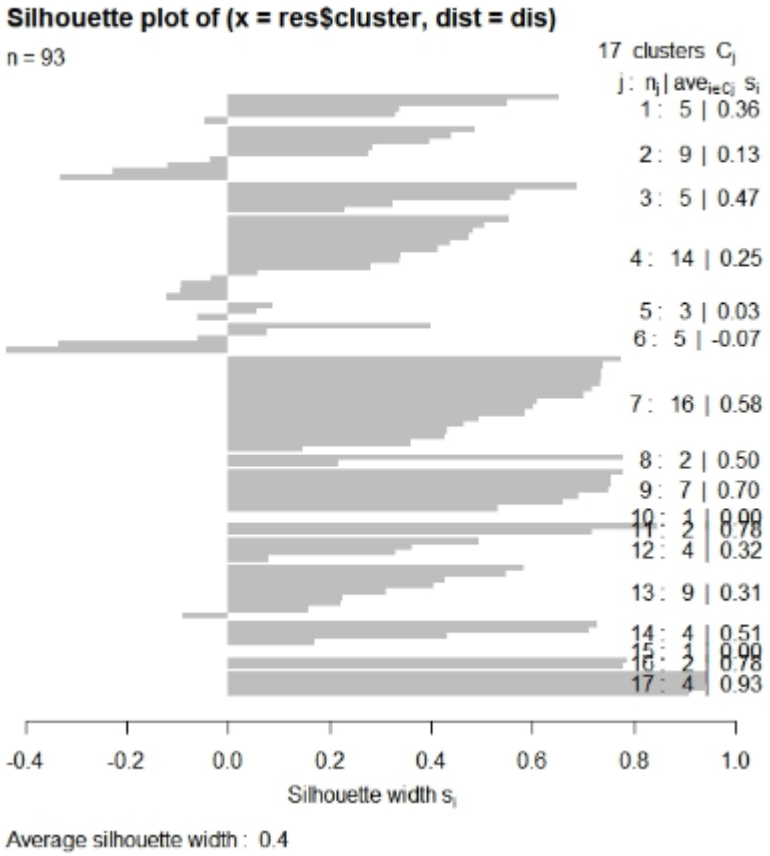


Average silhouette width : 0.38

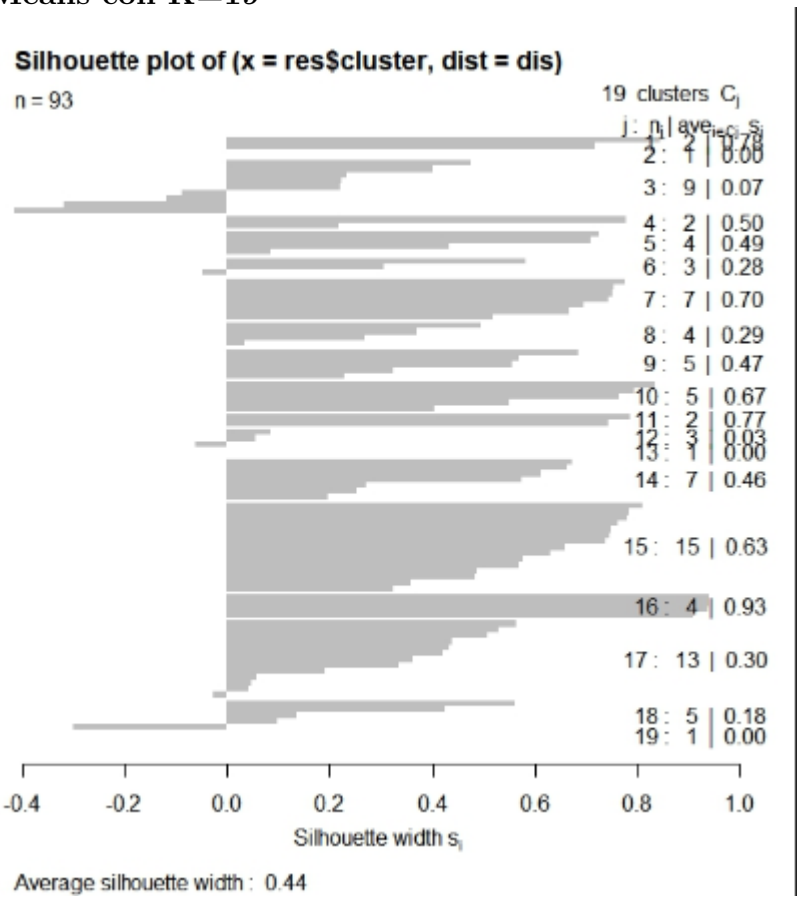
K-Means con K=15



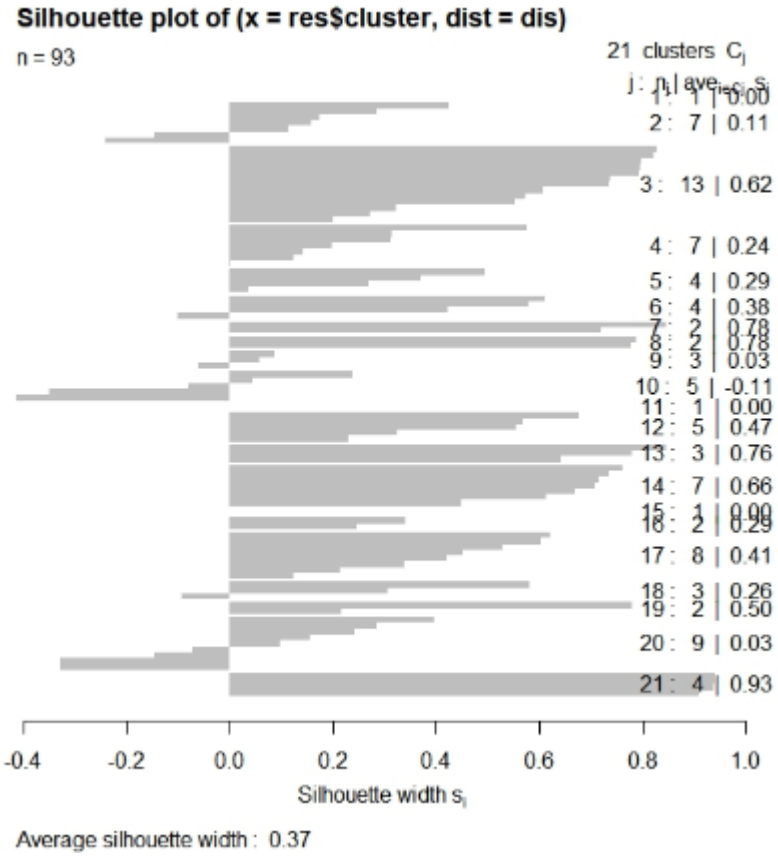
K-Means con K=17



K-Means con K=19



K-Means con K=21



16.3. Comparativa de tiempos

A continuación se muestra el detalle de los tiempos que emplean ambas alternativas de K-Means planteadas en el apartado 9. Se muestra el desglose por número de elementos analizado.

Iteración	Método 1 (seg)	Método 2 (seg)
1	0	0
2	0	0
3	0	0
4	0	0
5	0	0
6	0	0
7	0	0
8	0	0
9	0	0
10	0	0
Media	0	0

Cuadro 16: Rendimiento para 150 elementos

Iteración	Método 1 (seg)	Método 2 (seg)
1	0	1
2	0	1
3	0	1
4	0	0
5	1	1
6	0	1
7	0	0
8	0	0
9	0	0
10	0	0
Media	0,1	0,5

Cuadro 17: Rendimiento para 300 elementos

Iteración	Método 1 (seg)	Método 2 (seg)
1	0	0
2	0	1
3	1	1
4	0	0
5	0	0
6	0	0
7	0	1
8	1	1
9	0	0
10	0	1
Media	0,2	0,5

Cuadro 18: Rendimiento para 600 elementos

Iteración	Método 1 (seg)	Método 2 (seg)
1	0	1
2	1	0
3	0	2
4	0	0
5	1	0
6	0	1
7	0	1
8	0	0
9	1	1
10	0	1
Media	0,3	0,7

Cuadro 19: Rendimiento para 1200 elementos

Iteración	Método 1 (seg)	Método 2 (seg)
1	1	1
2	0	2
3	0	1
4	1	2
5	0	2
6	0	1
7	0	1
8	1	1
9	1	1
10	1	1
Media	0,5	1,3

Cuadro 20: Rendimiento para 2400 elementos

Iteración	Método 1 (seg)	Método 2 (seg)
1	1	10
2	2	4
3	3	5
4	2	5
5	3	4
6	4	7
7	4	4
8	4	3
9	3	6
10	4	3
Media	3	5,1

Cuadro 21: Rendimiento para 4800 elementos